electronics

MDPI

Article

Evolution of the "4-D Approach" to Dynamic Vision for Vehicles

Ernst Dieter Dickmanns

Independent Researcher, 85649 Brunnthal, Germany; edd@dyna-vision.de

Abstract: Spatiotemporal models for the 3-D shape and motion of objects allowed large progress in the 1980s in visual perception of moving objects observed from a moving platform. Despite the successes demonstrated with several vehicles, the "4-D approach" has not been accepted generally. Its advantage is that only the last image of the sequence needs to be analyzed in detail to allow the full state vectors of moving objects, including their velocity components, to be reconstructed by the feedback of prediction errors. The vehicle carrying the cameras can, thus, together with conventional measurements, directly create a visualization of the situation encountered. In 1994, at the final demonstration of the project PROMETHEUS, two sedan vehicles using this approach were the only ones worldwide capable of driving autonomously in standard heavy traffic on three-lane Autoroutes near Paris at speeds up to 130 km/h (convoy driving, lane changes, passing). Up to ten vehicles nearby could be perceived. In this paper, the three-layer architecture of the perception system is reviewed. At the end of the 1990s, the system evolved from mere recognition of objects in motion, to understanding complex dynamic scenes by developing behavioral capabilities, like fast saccadic changes in the gaze direction for flexible concentration on objects of interest. By analyzing motion of objects over time, the situation for decision making was assessed. In the third-generation system "EMS-vision" behavioral capabilities of agents were represented on an abstract level for characterizing their potential behaviors. These maneuvers form an additional knowledge base. The system has proven capable of driving in networks of minor roads, including off-road sections, with avoidance of negative obstacles (ditches). Results are shown for road vehicle guidance. Potential transitions to a robot mind and to the now-favored CNN are touched on.

Keywords: artificial Intelligence; spatiotemporal models; vehicle guidance

Citation: Dickmanns, E.D. Evolution of the "4-D Approach" to Dynamic Vision for Vehicles. *Electronics* **2024**, *13*, 4133. https://doi.org/10.3390/electronics13204133

Academic Editors: Sergio Trilles Oliver, Marco Del-Coco, Pierluigi Carcagni, Ditsuhi Iskandaryan and Xin Wang

Received: 20 August 2024
Revised: 14 October 2024
Accepted: 15 October 2024
Published: 21 October 2024

1. Introduction

The development of digital microprocessors (μPs) started in the 1970s; since then, growth in performance of about one order of magnitude every 4 to 5 years has been observed. The volume and power needed for a computer system stayed about the same, so that the system could fit into a (ground) vehicle. Studying computer vision for guidance of ground vehicles started in the 1960s [1] in the USA; the chapter cited gives a brief history of early activities in the field. When the author in 1975 received a call to a newly founded university in Munich, he decided to build a "Hardware-In-the-Loop" (HIL) simulation laboratory for developing the sense of vision for vehicles in general. This unusual step has paid off in the next decades. The first PhD thesis on vision for a road vehicle with this simulation loop appeared in 1982. Details on the 4-D approach may be found in [2–5]. In 1984, the first real test vehicle, a 5-ton van, was purchased and equipped as a test vehicle for autonomous mobility and computer vision: VaMoRs (Figure 1). In 1987, it drove fully autonomously on a free stretch of the new Autobahn A94 near Dingolfing with speeds up to the maximum of the vehicle: 96 km/h. After this demonstration, computer vision was accepted for both longitudinal and lateral control in the EUREKA-project PROMETHEUS from 1987 until 1994, replacing electromagnetic fields from buried cables for lateral guidance. The underlying differential equations forming the core of the new method for real-time visual perception based on feedback of prediction errors (Extended

Kalman Filters, EKF) for the point "here and now" were widely unknown in the computer vision community of that time.

Figure 1. Test vehicle 5-t van VaMoRs 1986 of UniBwM.

After the successful midterm 1991 demonstration in Torino with van-type vehicles, the up-to-then skeptical top management level of the car manufacturing company Daimler-Benz AG (DBAG) asked for a system in a passenger car for the final demo in 1994 near Paris, with the request to have passengers onboard the vehicle. We were willing to try this if twenty researchers could be funded by the project. To our surprise, the project was granted, and two Mercedes 500-SELs were selected as test vehicles, one each for DBAG and UniBwM. DBAG took care of all mechanical changes necessary in both vehicles, and UniBwM developed the vision system and all software necessary for autonomous driving with the new "Transputer"-system consisting of up to sixty processors. Figure 2 shows a survey of the UniBwM-system VaMP (short for VaMoRs-PKW). Both vehicles were the only ones capable of driving autonomously in 1994 in public three-lane traffic at speeds up to the maximum speed allowed in France of 130 km/h. Free driving, lane changing (see: https://dyna-vision.de/wp-content/uploads/2021/03/1994-Twofold-LaneChange-Paris-VaMP.mp4 (accessed on 14 October 2024)), and convoy-driving were demonstrated [5–9]. The structure of the second-generation vision system is shown in Figure 3. On the right-hand side, it shows the three levels with separate knowledge bases, as follows: gray: for features and generation of object hypotheses; green: for objects and subjects with the introduction of time by the 4-D approach; and red: for situation assessment and mission performance. Up to ten other vehicles could be detected and tracked in the own and the two neighboring lanes [10]. In total, more than 1000 km were driven autonomously.

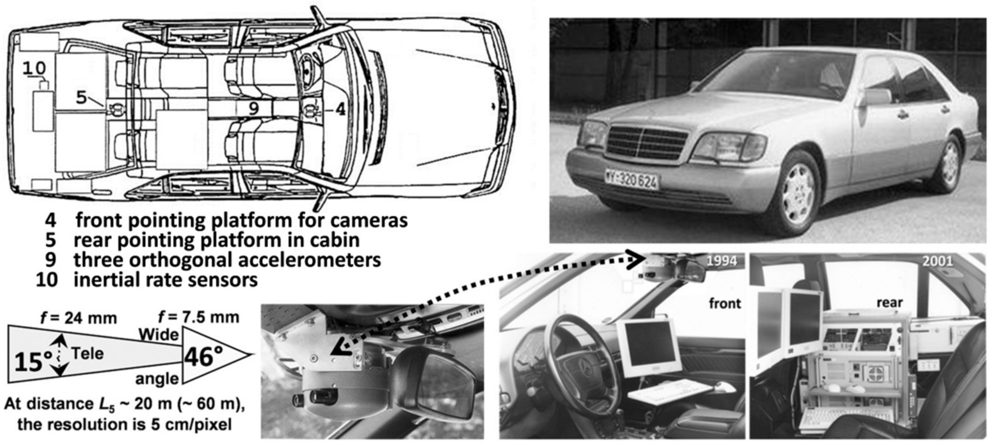


Figure 2. VaMP 1994: (**top left**) components for autonomous driving; (**right**) VaMP, a Mercedes sedan 500-SEL, and view into the passenger cabin (**bottom**); (**lower left**) bifocal camera arrangement (**front**) on a yaw platform.

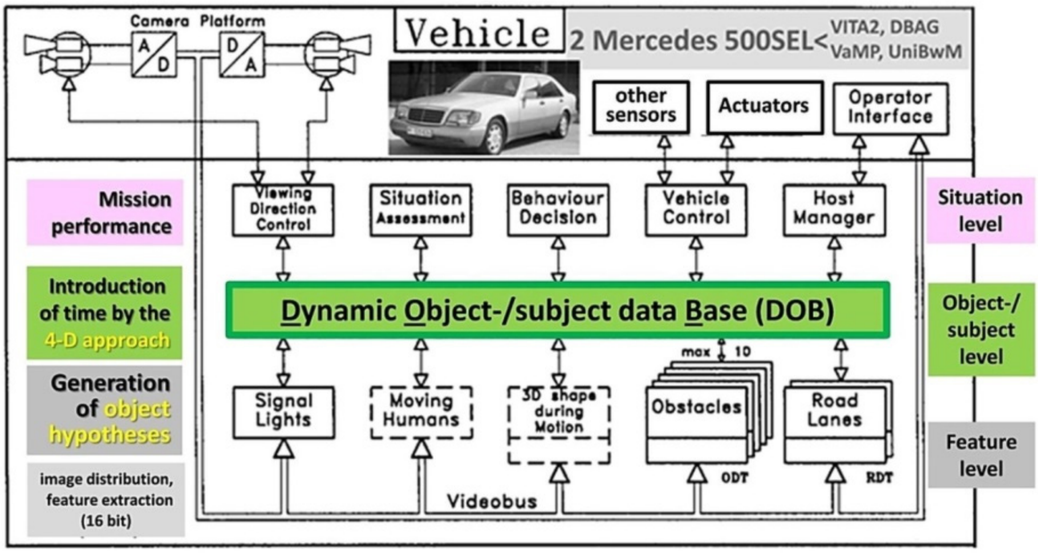


Figure 3. Second-generation vision system with up to 60 "transputers": real-time processing of four video-fields (**top left**) according to the 4-D approach. Three levels: The dynamic object data base (DOB in green) reduces the data volume on the feature level for the upper situation level (Situation Assessment for Behavior Decision and Control of Gaze and Attention (BDGA), as well as for Locomotion (BDL)) by two to three orders of magnitude without loss of essential information.

In the following year, 1995, the transputer system was replaced by more modern PC processors with ten times the computing power. This then allowed running the system at a full video rate (40 ms cycle time instead of 80 before) on only one-fifth the number of processors. With forward-looking cameras only, the fully autonomous long-distance test drive of Munich—Odense—Munich was performed in November 1995 [10,11] (Section 9.4.2.5).

At the end of 1996, the cooperation with DBAG was ended since a new project in cooperation with USA partners in the framework of an existing Memorandum of Understanding between the Departments of Defense was launched. The goal of the joint project AutoNav was to develop a next-generation vision system based on the 4-D approach capable of driving in networks of minor roads with sequences off-road; in addition to obstacles above the driving plane, negative obstacles (ditches) should also be detected and avoided autonomously.

The development of the PC market had advanced in the meantime, so that standard systems allowed building real-time vision systems by creating new software systems only. What should be the essential characteristics of our third-generation vision system?

2. From Local 4-D to Extended Maneuvers and Missions

The systematic exploitation of characteristics over time as the fourth dimension was the first goal: the local relations within the dotted yellow rectangle (upper left in Figure 4) are exploited in the 4-D approach through spatiotemporal models (differential equations) for feedback of prediction errors in order to adjust both state variables of the objects observed and parameters in the models used. The mission to be performed (in the lower right corner) is considered as a sequence of maneuvers, each maybe consisting of several maneuver elements. Capabilities for executing these maneuvers and missions can only be achieved by special time histories of control variables available in the real system. As far as the own body of the acting subject is concerned, these relations are part of a special knowledge base the subject has to learn. Since this action for control of motion is quite separate from perception, a special knowledge base for control of motion has been selected. It does not have to store the full trajectories of state variables but may be confined to just the time histories of the control variables involved, leading to the desired final state.

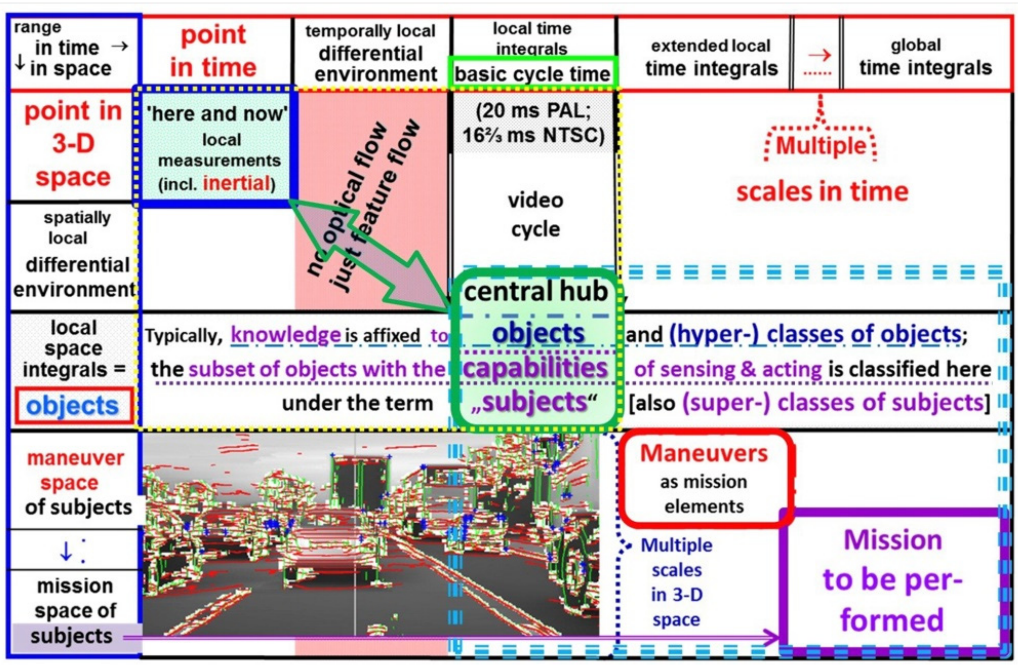


Figure 4. Spatial (vertical) and temporal (horizontal) structuring of the 4-D action space for multiple parallel visual/inertial perception.

In Figure 4, the dotted blue rectangle in the lower right shows this extension in space and time by the diagonal from the center to the bottom right. In the lower left corner, a road scene with dense traffic is depicted, not by an image consisting of pixels but by just five types of features yielding the same impression, as follows: white: image regions with nonlinear intensity changes in orthogonal directions; inclined edges found by vertical (red) and horizontal (green) search with special edge models according to [11] (chapter 5.2) and [12]; blue crosses for corners in intensity; and in gray: linearly changing image intensity values [13]. For a human observer of this artificial image of a scene, a correct interpretation is immediate. Even the number of several relevant objects on the three-lane road, with their approximate distances, will be recognized starting from nearby in the lowest row at the bottom of the image.

3. Why Multi-Focal Sets of Cameras

Due to perspective projection and the gaze direction being almost parallel to the ground, a pair of parallel lines on a planar ground (representing an idealized road) with the camera at the center of the near end is mapped into a triangle in the image with the tip at the far end (above the ground line nearby). The field of view of a sequence of single rows in the image of a camera also is a triangle, but with the tip in the camera. Since light rays are straight, a pixel in the image that transversally covers 1 cm at a distance of 10 m will cover 0.2 m at 200 m and even 2 m at a 2 km range. A vehicle of 2 m width will be covered by 200 pixels in one line at a range of 10 m and by ten pixels at 200 m; at a distance of 2 km, the width of such a vehicle is covered by a single pixel, so that the features of one vehicle are averaged away. This clearly shows that for understanding images of extended outdoor scenes, different resolutions should be used for imaging parts of the scene depending on the distance imaged and analyzed. This fact calls for multi-focal sets of cameras, at best with active control of the gaze direction (see [14]).

Experience in the past with a bifocal set of cameras has shown that in order to efficiently find the same region in the two images, the ratio in focal lengths should not exceed one order of magnitude. A factor of 6 to 7 has been found a good compromise, but the smaller the ratio, the easier it is to find the precise correspondence between image points. A "vehicle-eye" with four cameras according to Figure 5 has been proposed in [14]. Since one American partner in the AutoNav project contributed a new stereo vision system with two parallel-looking cameras, the configuration tested earlier looked like that shown in Figure 6, with the parallel stereo pair above the divergent-looking wide-angle cameras that may also be used for divergent stereo vision. At the center are the cameras with a standard field of view (FoV) and with high resolution.

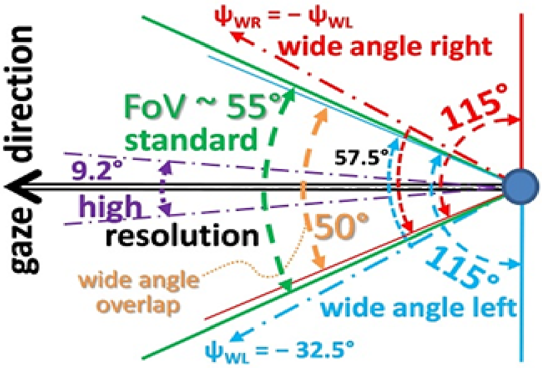


Figure 5. "Vehicle-eye" with three different focal lengths and two wide-angle cameras with divergent gaze direction [14].

Figure 6. Camera set used in VaMoRs in the AutoNav-project 1997–2003 with two pairs of stereo-cameras (parallel, outside top and divergent looking, bottom).

The size of a modern "vehicle eye" would be one to two orders of magnitude smaller than the test set shown here (see e.g., handy-cameras available actually). The request in resolution for the high-resolution camera of such an eye is that it should be able to make printed text readable with an acuity of edge localization of about 0.2 mrad/pixel (slightly better than human performance). With 800 pixels per row, the image then laterally covers an angle of about 9.2°. At 10 m distance, the lateral range covered by the camera is ~2.6 m with coverage of ~3.3 mm/pixel; at 200 m distance, the lateral range covered in the image will be ~50 m with 2 pixel/0.13 m. This is sufficient for recognizing the lane boundaries or the road limits marked by bright lines of, say, 0.12 m width. A divided highway with two lanes in each direction and one parking lane to the side may have a width of 20 to 30 m; so, at 200 m distance, just about twice the width of the highway is covered with high resolution. According to Table 1 (row 3), the total spread in resolution is 12; the lateral fields of view in degrees are 9.2 for high, 55 for medium, and 115 for low resolution selected. At a road crossing, the wide-angle camera should yield information on both roads intersecting at an approximately right angle.

In [14], a vision sensor covering the entire environment at different resolutions but mounted fixed onto the body of the vehicle has also been discussed.

For vehicles experiencing strong angular perturbations, like those from driving on rough ground, the images will be blurred under poor lighting conditions. Active control of the gaze direction allows stabilizing the gaze by feedback of rotational rates measured inertially by a set of inexpensive sensors directly on the platform of the eye. Figure 7 shows experimental results with VaMoRs for a standard braking maneuver. The reduction in amplitude for the gaze direction of the cameras (top curve) is more than one order of magnitude relative to the vehicle body (lower curve). Especially for interpreting the images of the high-resolution camera, this is an enormous alleviation when tracking a vehicle far away.

Table 1. Parameters of the cameras in the 'vehicle eye'.

Type (Resolution)	Low	Medium	High	Remark
Field Fields of view (in °)e	115 × 62	55 × 31	9.2 × 9.2	Left (−) and right (+) for 'low'
Imaging characteristics (resolution)	2.4 ¼ of med.	0.6 ⅓ of high	0.2	mrad/pixel, acuity of edge localization
Pixel/line	800	1600	800	These are rough estimates according to a pinhole model
Number of lines	450	900	800	
Data volume/frame	2.16 MB	4.32 MB	1.92 MB	3 Bytes/pixel; sum = 8.4 MB/cycle

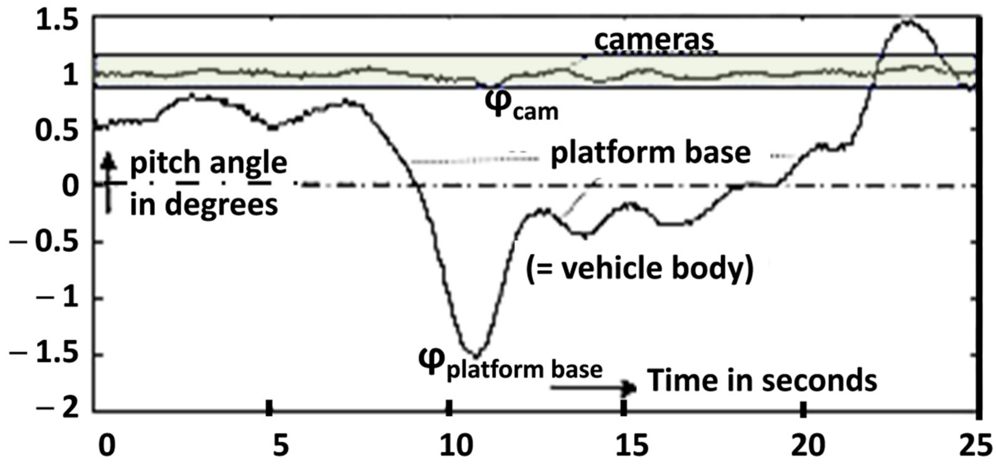


Figure 7. Gaze stabilization during a braking maneuver of VaMoRs. Reduction in the amplitudes of the cameras in pitch angle by more than one order of magnitude.

4. Efficiency Calls for Saccadic Gaze Control

A factor of 12 in resolution in both coordinates of the image leads to an increase in the number of pixels of 144 times for the high-resolution image as compared to low resolution. If the high-resolution cameras would be able to simultaneously deliver images of two pyramid stages (each yielding a reduction of ¼ additional pixels), that is (1 + 1/4 + 1/16) = 1.3125 times 144, this yields 189 times the number of low-resolution pixels in total on three image levels. For a total field of view of 360° × 45°, this results in ~212 Giga-pixels per video cycle. Beside the large number of high-resolution cameras needed for covering a 360° × 45° field of view (about 32 with ~2400 pixel per row and column), the total number of images per video cycle would increase to 52 {32 of 5.76 Giga-pixels plus (8 × 2 = 16) on the first pyramid level (with 1200 × 600 = 0.72 Giga-pixels for medium resolution) and 4 on the second pyramid level (with 600 × 600 = 0.36 Giga-pixels for low resolution)}; so the total number of image points would be (32 × 5.76 + 16 × 0.72 + 4 × 0.36) = 196.96 Giga-pixels, where, on the second pyramid level, two images in the same azimuth direction have been merged.

With two gaze-controllable eyes according to Figure 5 and Table 1, the number of cameras is reduced to eight (=25%), each with a much smaller image size and only one image per video cycle and camera (a reduction to 15%). The total number of image points is 2.8 Giga-pixels, corresponding to 1.4% of the sensor data from the high-resolution cameras mounted fixed onto the vehicle body. The factor of about 70 in data volume strongly favors the gaze-controlled vehicle-eye in addition to the reduced locations needed for mounting them around the vehicle. In [14], the two locations at the top end of the A-frame have been proposed as a promising compromise from several points of view. In particular, the tracking of traffic signs even up to a close approach in the second or third lane is easily handled. Experience has shown that tracking traffic lights and traffic signs at the side of the road and about 2 m above the ground clearly favors gaze control during approach. The black curve 2 in the right of Figure 8 shows two saccades of about 20° amplitude realized within 0.04 s each. Nowadays, the gaze direction could easily be controlled by locking in onto features of the sign, so that the green curve around 0.6 s would cover the ψ = 0 line. As can be seen from the video during the saccades, the blurred images of the scene cannot be used; during this short period, the vehicle has to live with the dynamical models and the resulting predictions. It takes about eight video cycles (20 to 30 msec) until the images can be interpreted correctly again. This corresponds to the delay time we humans also notice in our biological vision system [15].

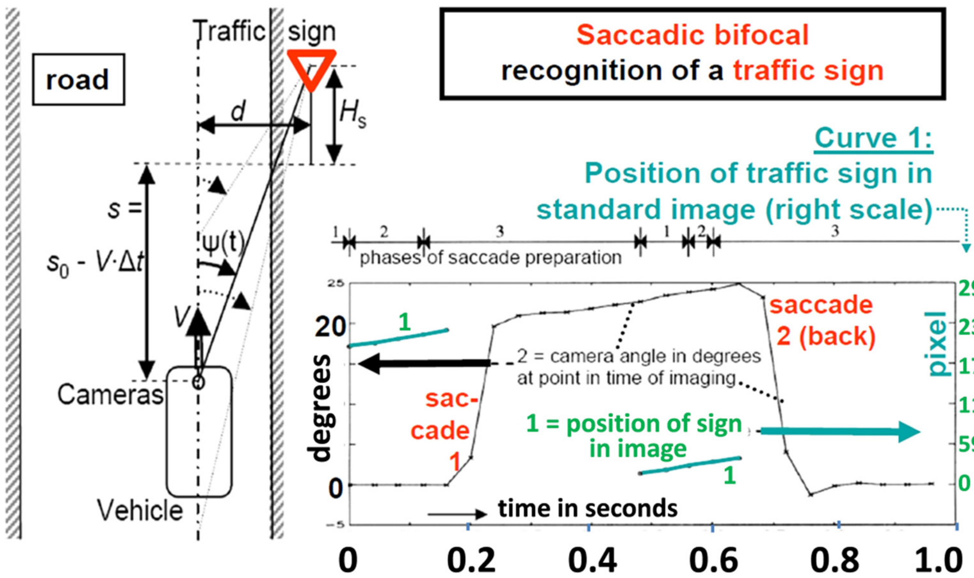


Figure 8. Detection and tracking of a traffic sign by saccadic vision in 1994 with VaMoRs. During this approach, both pan and tilt angles may become rather large. Curve 2 shows the gaze angle in degrees relative to the vehicle (see video https://dyna-vision.de/survey-of-experimental-results/table-of-firsts-in-dynamic-machine-vision-visual-guidance-of-autonomous-vehicles/ (accessed on 14 October 2024)).

Once gaze control is available in the "vehicle eye", it also allows precise tracking of special objects so that the high-resolution images are easier to interpret. With two independently gaze-controllable eyes, two single objects may be tracked and analyzed in parallel.

5. Three Levels of Scene Interpretation

5.1. Structuring the Task Domain on Temporal and Spatial Scales

An experienced human looking at the image in the lower left corner of Figure 4 cannot but recognize a road scene with at least three lanes; several types of vehicles drive in these lanes (represented in the rounded green square at the center). Bottom-up feature extraction and data evaluation at the point "here and now" are done for the entire last image (upper-left corner). The results are communicated to the central box in Figure 4 running at the video rate; this central unit tracks the hypotheses for objects and subjects in 3-D space over time. The results are stored in a dynamic object data base (DOB) using a scene tree (see Figure 9, white field within the green rectangle) as a communication device to the situation level 3. On level 2, time is introduced, allowing temporally deeper scene understanding with respect to maneuvers and to the overall mission.

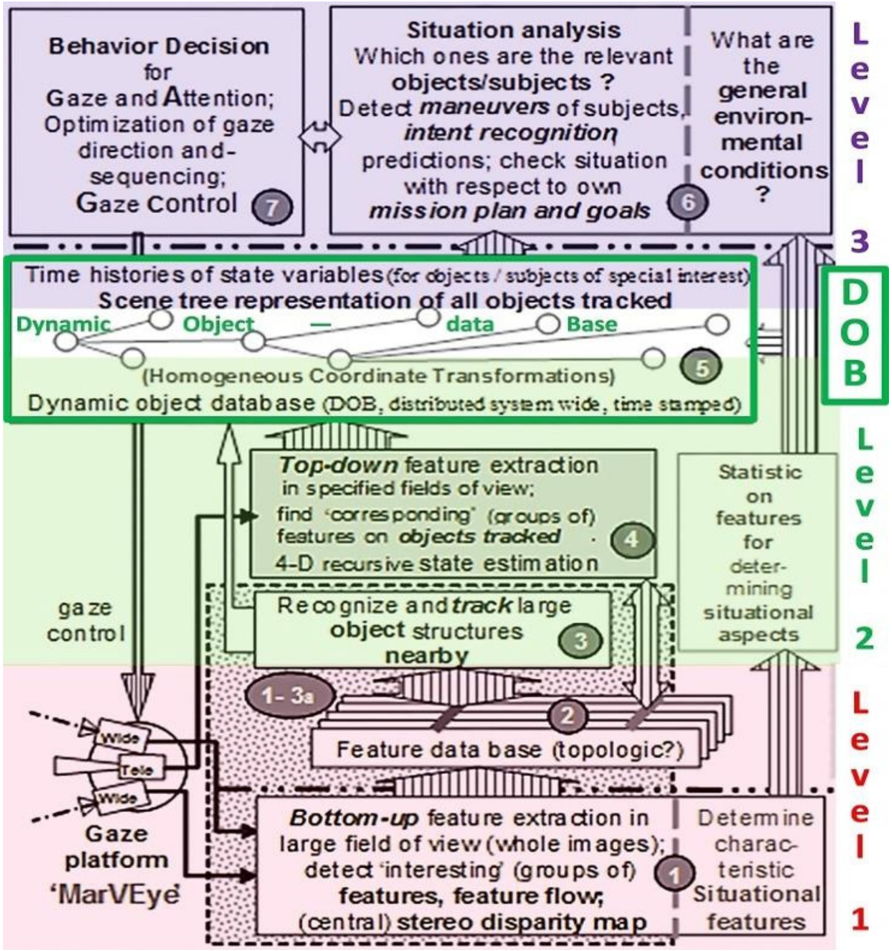


Figure 9. Three levels for dynamic scene understanding in the 4-D approach. Red: measurement data and image features; green: object data and pose estimation; magenta: situation assessment and mission data.

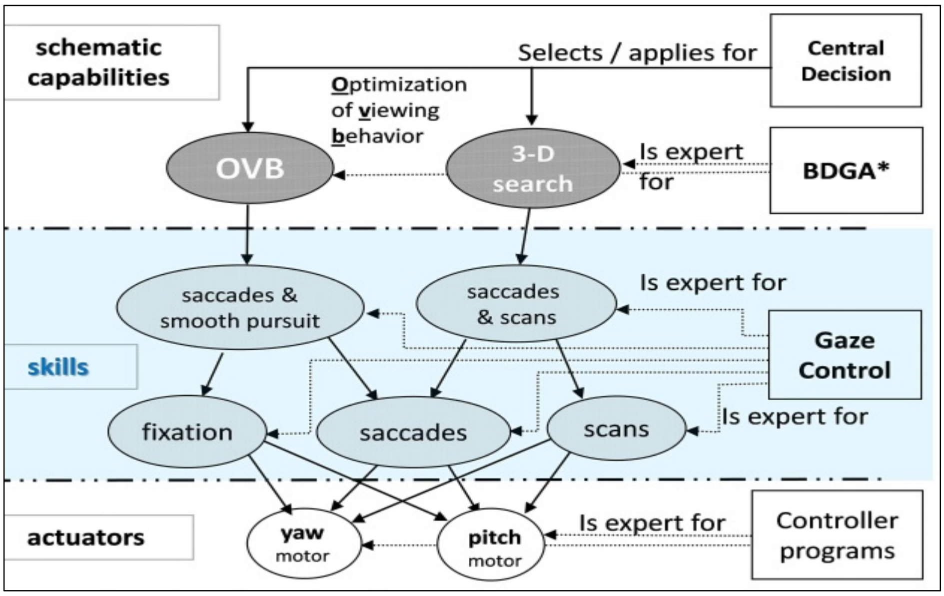
Three levels have been implemented for visual cognition and task performance; Figures 4 and 9 indicate that an organization of the overall process of performing a mission autonomously should be structured according to three levels:

1. Image features and other sensor data (bottom-up in each cycle);
2. Objects (objects proper and subjects as separate classes) in 3-D space and time (4-D) {in green};
3. Mission performance with a special knowledge base on maneuvers: missions consist of a consecutive list of mission elements which are built by a sequence of maneuvers and their elements (top, magenta).

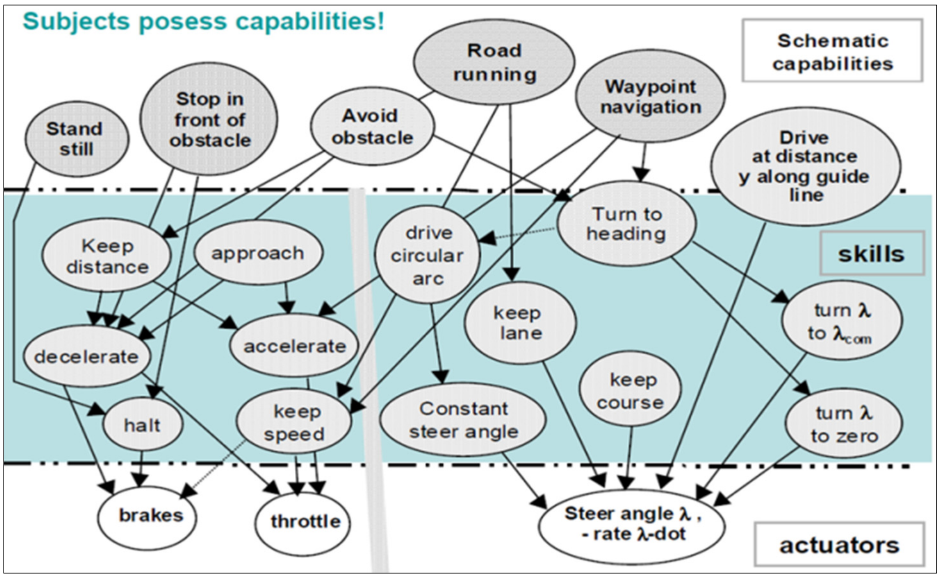
Level 1 is mainly bottom-up and has to deal with large amounts of sensor data. Levels 2 and 3 may interact every now and then with level 1 with respect to looking for special features derived from knowledge on classes of objects and on mission elements. Level 3 represents the best adaptations of the internal mental visualizations to the external real world by 3-D spatial models of objects and of temporal processes (4-D); this is the knowledge base for perceiving the semantics of the outside world (derived from the sensor data from levels 1 and 2) and for efficient performance of the mission.

The result is a set of capabilities both for perception and for mission performance; each of these capabilities achieved by a subject is called a skill it has. There are capabilities for behavior decision in gaze and attention (top row in Figure 10a), which link the mental decisions for gaze control to the actual hardware for realizing them in the subject (bottom row). This shows the capability network for Behavior Decision for Gaze and Attention

(BDGA); the group of software that leads to certain skills for realizing the capabilities is marked by a blue background. The arrows indicate which subsystems are involved in realizing the behavior. Figure 10b shows the capability networks for Behavior Decision in Locomotion of ground vehicles (BDL). Again, the capabilities are shown in the top row, and the actuators available for realizing them are seen at the bottom. Before activation of a skill, it is checked whether all subsystems needed are actually available; this is important for autonomous detection of errors that might have occurred in the meantime. Figure 10a,b show the capabilities as a flexible concept for linking the mental world to real-world hardware in order to perform maneuvers and missions.



(**a**)



(**b**)

Figure 10. (**a**) Behavior Decision for Gaze and Attention (BDGA). Bottom: activation of hardware; center: learned skills; top: tested and proven knowledge about capabilities of a subject. (**b**) Behavior Decision for Locomotion of ground vehicles (BDL). Left-hand side: longitudinal control; right-hand side: lateral control.

The transition from mental decisions in the main computer system to their realization in the real world with the actuators of the subject is done by a dual representation:

1. With AI-methods on the mental side with extended state charts containing the conditions for transition between the modes [16];
2. With methods from systems dynamics for realization on (embedded, distributed) processors close to the actuators by feed-forward and feedback control laws [17] (see Figure 11).

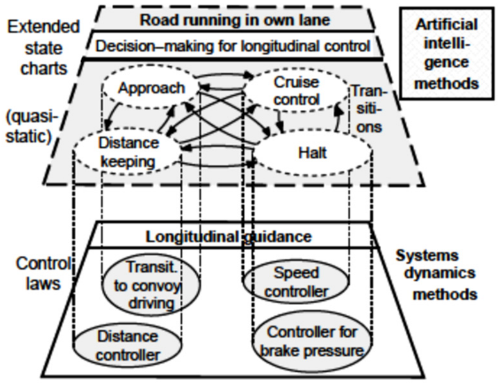


Figure 11. Dual representation of behavioral modes with methods from artificial intelligence (top, dashed) and systems dynamics (bottom).

Details on the realization may be found in [16–21], summarizing the results of the AutoNav project until that year, and in [22–24]. A video on the final demo may be viewed under https://dyna-vision.de/wp-content/uploads/2021/03/2003%20EMS-On-OffroadDriving%20VaMoRs.mp4 (accessed on 14 October 2024).

5.2. Multiple Parallel Feedback Loops in Perception and in Control of Behavior

Many parallel feedback loops result in an overall system for perception of the actual situation and for autonomous execution of a mission. Figure 12 presents a survey on the third-generation system of UniBwM for perception and control of a mission. The mental part, encompassing values and goals in decision making, has become dominant (see top of the figure). The subject now is no longer just part of the material evolution, but starts understanding at least part of the observed processes of evolution. Beside its own mental model for the processes observed, it has its own values and goals, and in general it will try to move towards an improved state with respect to its individual feelings and its thoughts about the mission. Note that the individual and the cultural value systems of the group may diverge in several points of view, yielding potential conflicts in decision making.

At the bottom of Figure 12 is the material body of the subject with all its sensors and actuators. State prediction to the next point of sensor data evaluation and control output is the central activity of the mind (upper-right corner of the figure). This triggers both the feed-forward control for maneuvers performed and feedback components for minimizing prediction errors due to unforeseen perturbations from the environment (wide downward arrow on the right-hand side). From the assessment of the actual situation (in the upper-left corner) the best direction for visual perception by multi-focal vision is derived and communicated to the corresponding actuators for the control of viewing direction (pink arrow). Another set of feedback loops is included for tracking the features of all objects of interest (black arrow).

The "Gestalt"-idea of spatial objects mapped from 3-D to 2-D by perspective projection assists in deriving hypotheses for objects discovered from collections of features in the images (green arrow). For validated hypotheses of objects, the blue arrows support efficient tracking of objects over time. The broad horizontal arrow forms the basis for detecting effects of perturbations on the vehicle carrying the sensors. Finally, temporal predictions allow changes in the given situation to be expected (dark green arrow). The mental aspects of all these loops have thus become predominant in control of perception and behavior. Note that the knowledge bases for the three levels feature the following:

1. Extraction and grouping of image features for the step following;
2. Generation of object hypotheses and their tracking over time;
3. Situation assessment, including derivation of control for mission performance, which has to be supported by special interconnected software systems representing the foundation of skills that link the mental world to applications in the real world.

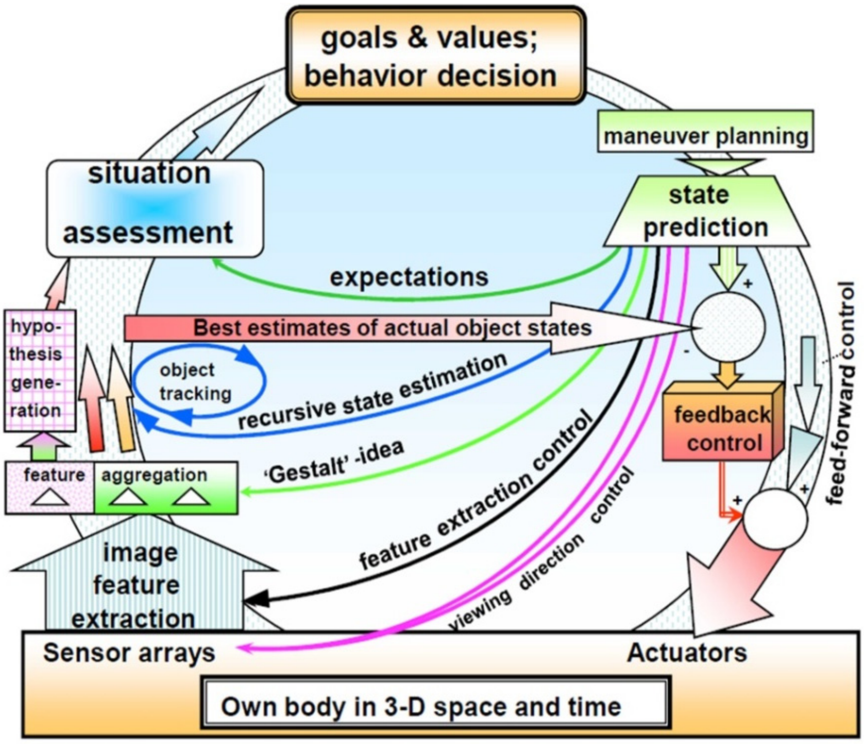


Figure 12. Multiple feedback loops of the 4-D approach for autonomous performance of a mission form the core of the "robot mind".

Figure 13 sketches the overall resulting system. The stereotypical capabilities shown on the horizontal connection between BDGA (red rectangle on the left side) and BDL (blue rectangle on the right) in the vertical center of the gray-shaded region now constitute the core of the autonomous system. The upper part shows, by the arrows, the capabilities that are needed for each task (road/lane following, turning off onto a crossroad, following a sequence of waypoints). The lower part indicates the basic skills that are needed for each capability. The dashed block at the left-hand side in Figure 13 indicates that the mission to be performed (mission plans, lower right corner in Figure 4) and the behavioral capabilities actually available have to be provided by the human initiator of the mission. More details may be found in the set of publications at the International Symposium on Intelligent Vehicles [16–21] in Dearborn USA. The large rectangle shaded in gray contains all behavioral capabilities available, the stereotypical capabilities to realize them (between BDGA and BDL) in the center, and the skills realized with the subsystems (small circles in the lower part). The lower small rectangles group these skills into three behavioral fields: gaze control (left), visual perception (center), and vehicle control (bottom right). Depending on the sensor systems and the control systems available in the subject, the set of capabilities and skills may be extended to perform other types of missions.

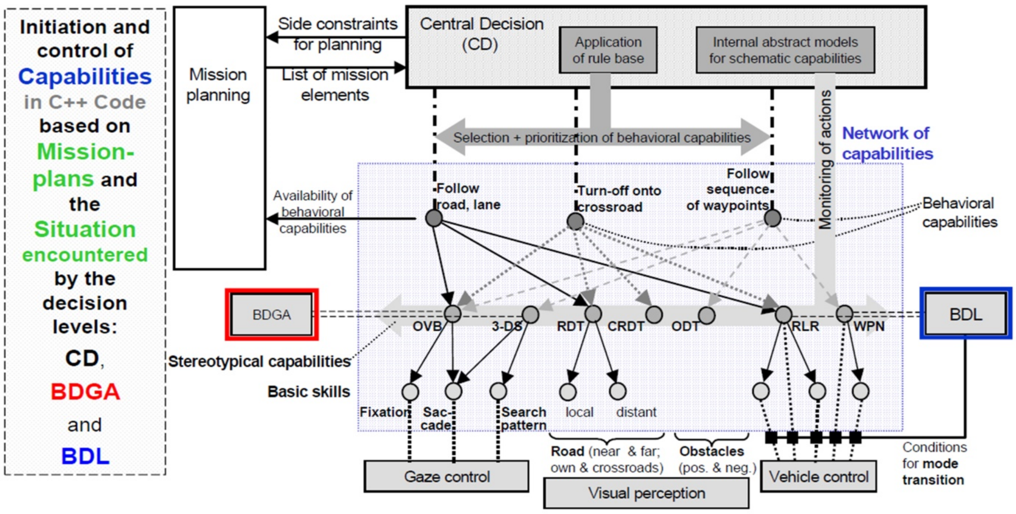


Figure 13. Behavior Decision for Gaze and Attention (BDGA in red) and for Locomotion (BDL in blue); abbreviations: OVB = Optimization of Viewing Behavior, 3-DS = 3-D Search, RDT = Road Detection and Tracking, CRDT = Cross Road Detection and Tracking, ODT = Object Detection and Tracking, RLR = Road and Lane Recognition, WPN = Way-Point Navigation.

6. From the 4-D Approach to Expectation-Based, Multi-Focal, Saccadic EMS-Vision

The 4-D approach at the core of EMS-vision allows generating around the point "here and now", by feedback of prediction errors, some kind of consciousness grounded on the internal knowledge bases and the adaptation of spatiotemporal models available in them. The strict distinction between state and control variables in these models helps to reduce storage requirements for extended maneuvers and missions. Time histories of control variables specify maneuver elements and maneuvers for achieving desired final values of the state variables. Unforeseeable perturbations during performance of maneuvers can immediately be counteracted by feedback of errors between the desired and the actually developing trajectory (see right-hand side of Figure 12). Figure 14 shows a typical result for a lane-change maneuver with the test vehicle VaMP. The green straight lines in the upper-left sub-image show the commanded constant steer rates between seconds 55 and 63. It can be seen that the actual steer rates (black curve) deviate quite a bit from the nominal green one. The other sub-images show the corresponding state variables, as follows: top right: the steer angle λ; bottom left: the yaw angle of the vehicle; and bottom right: the lateral offset with the switch to the new lane as reference at the center of the maneuver.

Since, beside the state variables in the models for the dynamic behavior of objects observed, parameters in the structure of these models may also be adapted, the overall system will look like that shown in the sketch in Figure 15. In parallel to the real world (at left), an internally represented imaginary world is constructed by feedback of prediction errors exploiting spatiotemporal (4-D) models (lower right part, "tracking"). Groups of features are assigned to hypothesized objects mapped by the laws of perspective projection. A number of objects, n (including subjects), are tracked in parallel (upper-left part of the "tracking"-circle in the lower right). Unassigned features lead to hypotheses of new objects ("detection", center upward), which are tested for a few cycles before a new object is added to the list for tracking. Parameters of the models used may also be adapted to reduce prediction errors ("learning" in center top). By storing new successful sets of parameters, the background knowledge is increased from experience (center top). The results of analyzing all observations together with the mission to be performed lead to the control output onto the real vehicle (top loop); these control variables are also fed into the models used for object recognition and their adaptation. This may be a first small step in the direction of a robot mind (see area shaded in gray in Figure 15).

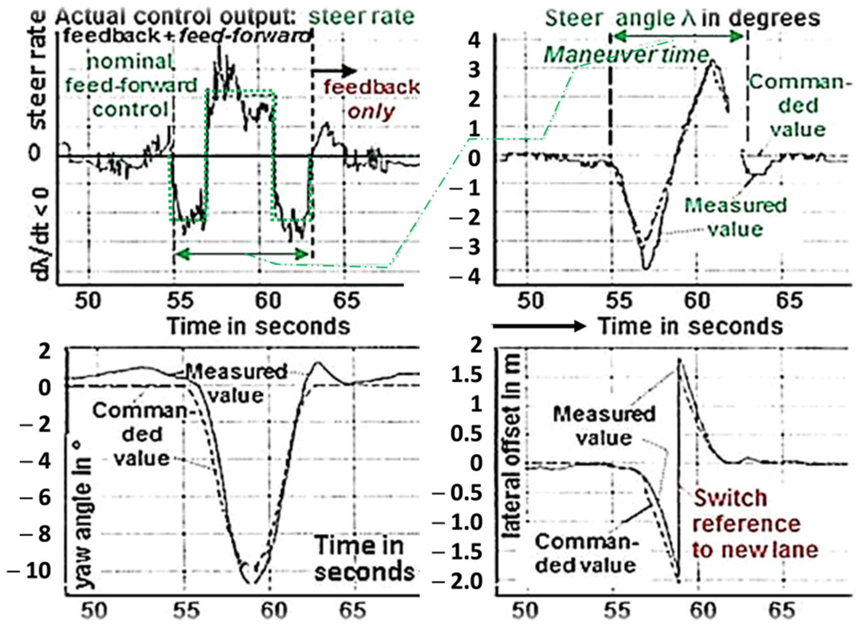


Figure 14. Maneuvers as knowledge elements for vision and control: real lane change maneuver with VaMP, 1994.

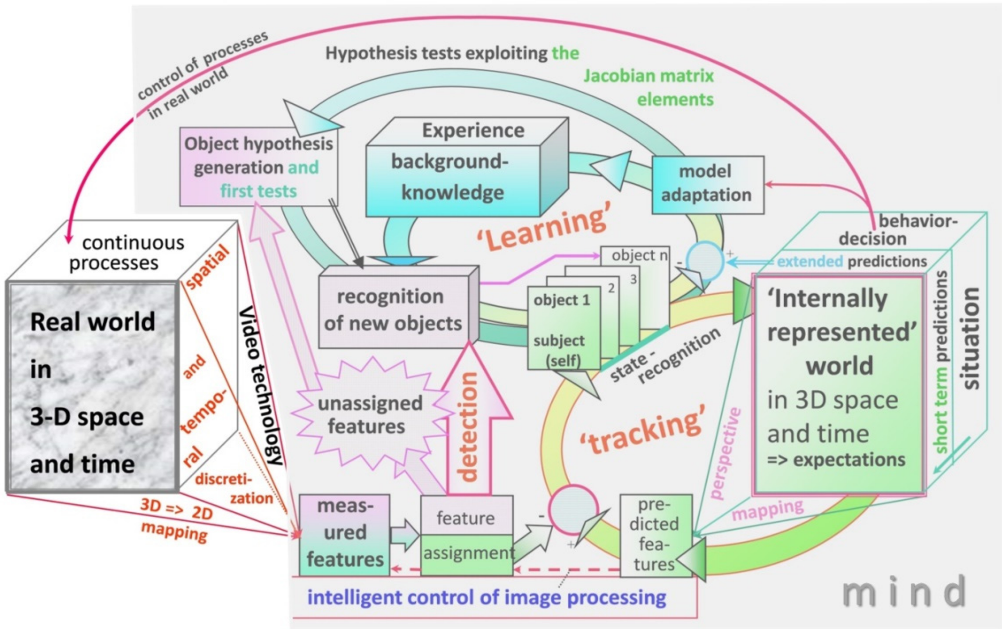


Figure 15. Generation of a mental world in parallel to the real outside world (rectangle at left) based on measurement values and the 4-D approach by feedback of prediction errors "here and now".

The complex robotic system that has evolved on the basis of the 4-D approach is hard to describe in single figures like Figures 13 and 15. The system realized in the project AutoNav has proven to be able to drive autonomously on a network of minor roads and off-road on hard ground, detecting and avoiding an unknown ditch. The video on the final demonstration with VaMoRs in 2003 may be viewed under https://dyna-vision.de/wp-content/uploads/2021/03/2003%20EMS-On-OffroadDriving%20VaMoRs.mp4 (accessed on 14 October 2024).

7. On the Way to a Robot Mind

In the long run, when several subjects observe the same scene by using similar dynamic models, these subjects may exchange their results in order to agree on the models best suited for the actual situation. The availability of a language including common terms for objects, subjects, values, and different actions is a prerequisite for developing a robot mind. A robotic subject with a mind should be able to distinguish between itself (the "I") and the rest of the world (summarized in Figure 16 within the dash-dotted ellipse). The five words (I think, therefore I am) printed in large bold letters are the once famous statement of Descartes, who claimed in 1637 the mind to be a separate substance; in the meantime, that became obsolete with Damasio [25].

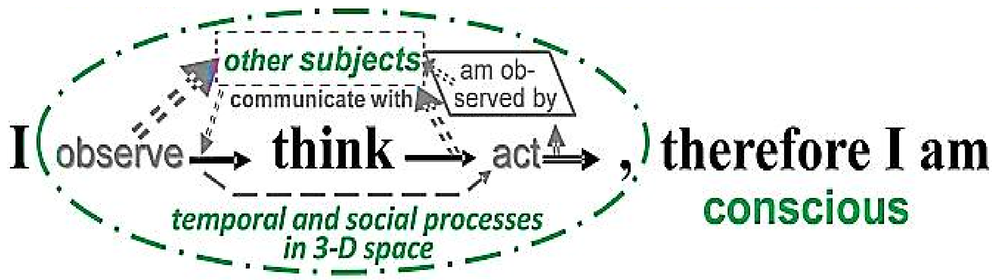


Figure 16. Visualization of closed loops needed between several subjects for the development of a common mind.

Today, the mind is even often considered not to be a separate subject but a new quality in the material world that emerged after biological subjects had developed the capabilities of sensing, storing, retrieving, and processing of data and knowledge, as well as decision making and acting towards certain goals. Mind emerged when these capabilities had reached a higher level of evolution [26–31]. In the meantime, very advanced robotic vehicles may also have these capabilities. Will they have a mind?

Humans tend to talk of an object (subject) having a mind if it can make decisions, increasing the value of an observed system or of improving its own state. A system realizing all feedback loops in Figure 12 certainly may be considered to have a mind. If goals and values for all potential points in the 4-D matrix of Figure 4 are essential for decision making, this over-arching concept may be called a mind. Of course, there will be many different levels of mind depending on the knowledge bases and the capabilities for both perception and for action of the subject. At present, many of the capabilities for communication and for adaptation of the dynamic models used are with the human operators of the systems; however, there is no reason why this could not also be implemented in robotic vehicles. An exception is the fear sometimes mentioned that these capabilities could make humans superfluous. Another approach to mind generated by technology may be found in [32–34] from our American partners in the AutoNav project.

8. General Development in the Field "Autonomous Driving"

A survey of the development of automatic driving and autonomous road vehicles is given in [35,36]. Computer vision has become a field of intensive research since the mid-1980s with the project "Strategic Computing" in the USA [37] and the European project "Prometheus" [38]. Developments in the USSR are summarized in [39], where actions of several agents (subjects) in spatial environments are considered essential for developing a mind.

Since 1992, an "International Symposium on Intelligent Vehicles" has been held, with locations changing regularly between three regions: USA, Europe, and Asia (abbreviated IV'xy, where xy stands for the last two numbers of the year). To obtain an overview of the development of intelligent vehicles, scanning the proceedings of the IV'xy is a good starting point. However, there continue to be relevant publications in conferences and journals, like SPIE-Mobile Robots, ICCV, ICPR, ICRA, IROS etc.

Even though the 4-D approach has been published regularly at various international conferences and workshops since 1987, its acceptance for own developments has been rather low. Our ten-year cooperation with Daimler-Benz AG from 1986 to 1996 has led, at least, to partial acceptance. The cooperation with the US American National Institute for Standards and Technology (NIST) from 1997 until 2003 in the project AutoNav resulted in the adoption of the name "4D" into their software system [33]. The successful demonstrations with the American test vehicles led to the announcement of the "Grand Challenge" by DARPA in 2002. However, since, in the meantime, the Global Positioning System GPS, funded by DARPA, had become functioning, the demo requested for the "Grand Challenge" was no longer required to find the path for vehicle guidance by vision with an onboard video system. The route was prescribed by a dense sequence of GPS waypoints handed to the participants a few minutes before the start.

Of course, this alleviated the vision task considerably, since the autonomous vehicle now had to just avoid obstacles above the driving plane. In order to detect obstacles reliably, DARPA had also funded the development of LIDAR-based obstacle detectors. This has led to the fact that one of the test vehicles for the demo in 2005 did not even have video cameras on board. The test results were widely publicized, so that for many readers and viewers, this was the beginning of autonomous driving. In the Urban Challenge of 2007, there were additional driving vehicles in the specially built village that was used as a test site. The routes to be driven were again prescribed by (now more widely separated) GPS waypoints.

After these events, the company Google generously installed a special institute for the winner of the 2005 Grand Challenge on its campus. This has led to the fact that the approach based on GPS-waypoints and high-precision maps, together with a 360° revolving laser-system, in addition to a few video cameras mounted fixed onto the body of the car, has become the standard approach in the civil market for autonomous driving. A few companies offering the capability of partially autonomous driving with their cars prefer to have laser range finders at the corners of the vehicle closer to the ground instead of on the roof. This type of vision on preselected routes for driving has been named "confirmation-type" vision [40], since characteristic stationary objects near the route with well-recognizable features are usually also indicated in the high-definition maps provided to the autonomous vehicle.

On the contrary, the 4-D approach taking the dimension of time directly into the models for visual perception leads to results with less delay time and to a strong reduction in the amount of data to be handled in parallel. This has been dubbed "scout-type" vision in [40]. The critical point here is to find an early transition from feature data in the image to 3-D objects moving in the outside world in real-time. A few video cycles have been shown to be sufficient for supporting hypotheses for interpretation, leading to a time delay for object recognition of a few tenths of a second (similar to human perception).

Computational neural net approaches were investigated early (e.g., [41]) but did not make it to market applications, due partially to the high computing power needed. This changed lately when deep neural nets profited from the µP development. In recent publications on convolutional neural networks (CNNs) and vision applications, one can find the topic "Machine vision applications that require real-time performance". Typical for these papers is that the computational neural nets work on several images of the video sequence simultaneously [42–53]. This involves large amounts of data and leads to increased time delay for controlling real-time processes. Many researchers now seem to also believe in CNNs for perceiving general road environments and for vehicle guidance. The future has to show the best approach available.

The advantage of the "4-D approach" is that due to the dynamical models involved, only the last image of the sequence needs to be worked on, both for state estimation of the real-world objects seen and for simultaneously adapting parameters in the models used. This eliminates the old arguments of the Greek philosophers Plato and Aristotle that either the sensor data or the mind has to be accepted as basic elements for perception and thinking. However, evolution of the mind in groups of communicating and cooperating individual subjects dealing with real-world challenges had not been considered. This is the achievement of the 4-D approach in robotics.

The group at UniBw Munich over the last two decades has concentrated on recognizing unstructured scenes with vision and laser range finders. They developed the capability of autonomous driving with real cars in this usually complex environment using the conventional approach [54–58]. Several car manufacturers already offer level-3 autonomous driving in some of their high-end cars; Daimler in September 2024 announced it for highway driving up to speeds of 60 km/h. Other international car makers have been more daring in the past, but had to reconsider their decision after accidents. The general attitude towards autonomous driving seems to have become more cautious since the potential scenes to be recognized may be quite varying.

9. Conclusions and Outlook

The 4-D approach to dynamic vision at the core of EMS-vision allows an immediate realization of the central part of consciousness for a robotic vehicle. It knows where it is relative to the road, and what type of objects and other subjects are relative to its own position. Including the capabilities of performing maneuvers based on sequences of simple maneuver elements and of counteracting perturbations experienced by feedback of prediction errors makes the method very efficient, including for environments that are hard to correctly predict.

It is interesting to note that feedback of prediction errors has recently become a topic in cognition in the fields of psychology and philosophy [59–63]. It would be interesting to check whether EMS-vision combined with neural net methods could merge the positive aspects of both approaches.

Funding: This research received no external funding since 2004.

Data Availability Statement: Data are contained within the article.

Acknowledgments: Much of the content and parts of some figures have been derived from contributions of Ph.D. students over three decades. Since the dissertations are in the German language, I have refrained from citing them. The interested reader may access them via the website www.dyna-vision.de (accessed on 14 October 2024) and the names in references [2,6–9,16–21]. The overall research was possible only by the generous support through UniBwM and the department for Aero-Space Technology (Luft- und Raumfahrt Technik LRT). My thanks go to all colleagues and leading offices. Daimler-Benz AG enabled us to become part of the Prometheus project and get the fastest vision-based autonomous cars, “Mercedes 500 SEL”, which allowed us to drive in 1994 near Paris in public three-lane French Autoroute traffic with speeds up to the maximally allowed 130 km/h.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Moravec, H. Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover. Ph.D. Thesis, Computer Science Department, Stanford University, Stanford, CA, USA, March 1980; Chapter 2: History.
2. Mysliwetz, B.; Dickmanns, E.D. A Vision System with Active Gaze Control for real-time Interpretation of Well Structured Dynamic Scenes. In Proceedings of the 1st Conference on Intelligent Autonomous Systems (IAS-1), Amsterdam, The Netherlands, 8–11 December 1986; Hertzberger, L.O., Ed.; pp. 477–483.
3. Dickmanns, E.D.; Graefe, V. Dynamic monocular machine vision. *Mach. Vis. Appl.* **1988**, *1*, 223–240. [CrossRef]
4. Dickmanns, E.D.; Graefe, V. Applications of dynamic monocular machine vision. *Mach. Vis. Appl.* **1988**, *1*, 241–261. [CrossRef]
5. Dickmanns, E.D. Historical development of use of dynamical models for the representation of knowledge about real world processes in machine vision. *Signal Process.* **1994**, *35*, 305–306. Available online: https://dyna-vision.de/home/introduction/temporal-development-of-dynamic-vision/ (accessed on 14 October 2024). [CrossRef]
6. Dickmanns, E.D.; Behringer, R.; Dickmanns, D.; Hildebrandt, T.; Maurer, M.; Thomanek, F.; Schiehlen, J. The Seeing Passenger Car ‘VaMoRs-P’. In Proceedings of the Intelligent Vehicles ’94 Symposium, Paris, France, 24–26 October 1994; pp. 68–73.
7. Schiehlen, J.; Dickmanns, E.D. A Camera Platform for Intelligent Vehicles. In Proceedings of the Intelligent Vehicles ’94 Symposium, Paris, France, 24–26 October 1994; pp. 393–398.

8. Thomanek, F.; Dickmanns, E.D.; Dickmanns, D. Multiple Object Recognition and Scene Interpretation for Autonomous Road Vehicle Guidance. In Proceedings of the Intelligent Vehicles '94 Symposium, Paris, France, 24–26 October 1994; pp. 231–236.
9. Behringer, R. Road recognition from Multifocal Vision. In Proceedings of the Intelligent Vehicles '94 Symposium, Paris, France, 24–26 October 1994.
10. Thomanek, F. Visuelle Erkennung und Zustandsschätzung von Mehreren Straßenfahrzeugen Zur Autonomen Fahrzeugführung. Ph.D. Thesis, Universität der Bundeswehr München, Neubiberg, Germany, 1996.
11. Dickmanns, E.D. *Dynamic Vision for Perception and Control of Motion*; Springer: Berlin/Heidelberg, Germany, 2007; 474p, Section 9.4.2.
12. Hubel, D.H.; Wiesel, T. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *J. Physiol.* **1962**, *160*, 106–154. [CrossRef] [PubMed]
13. Dickmanns, E.D.; Wuensche, H.-J. Nonplanarity and efficient multiple feature extraction. In Proceedings of the First International Conference on Computer Vision Theory and Applications (Visapp), Setubal, Portugal, 25–28 February 2006.
14. Dickmanns, E.D. May a pair of 'Eyes' be optimal for vehicles too? *Electronics* **2020**, *9*, 759. [CrossRef]
15. Pöppel, E.; Chen, L.; Glünder, H.; Mitzdorf, U.; Ruhnau, E.; Schill, K.; von Steinbüchel, N. Temporal and spatial constraints for mental modelling. In *Frontiers in Knowledge-Based Computing*; Bhatkar, V.P., Rege, K.M., Eds.; Narosa Publishing House: Delhi, India, 1991; pp. 57–69.
16. Gregor, R.; Lützeler, M.; Pellkofer, M.; Siedersberger, K.H.; Dickmanns, E.D. EMS-Vision: A Perceptual System for Autonomous Vehicles. In Proceedings of the IEEE Intelligent Vehicles Symposium 2000, Dearborn, MI, USA, 5 October 2000; pp. 52–57.
17. Pellkofer, M.; Dickmanns, E.D. EMS-Vision: Gaze Control in Autonomous Vehicles. In Proceedings of the IEEE Intelligent Vehicles Symposium 2000, Dearborn, MI, USA, 5 October 2000; pp. 296–301.
18. Siedersberger, K.-H.; Dickmanns, E.D. EMS-Vision: Enhanced Abilities for Locomotion. In Proceedings of the IEEE Intelligent Vehicles Symposium 2000, Dearborn, MI, USA, 5 October 2000; pp. 146–151.
19. Gregor, R.; Dickmanns, E.D. EMS-Vision: Mission Performance on Road Networks. In Proceedings of the IEEE Intelligent Vehicles Symposium 2000, Dearborn, MI, USA, 5 October 2000; pp. 140–145.
20. Hofmann, U.; Rieder, A.; Dickmanns, E.D. EMS-Vision: An Application to Intelligent Cruise Control for High Speed Roads. In Proceedings of the IEEE Intelligent Vehicles Symposium 2000, Dearborn, MI, USA, 5 October 2000; pp. 468–473.
21. Lützeler, M.; Dickmanns, E.D. EMS-Vision: Recognition of Intersections on Unmarked Road Networks. In Proceedings of the IEEE Intelligent Vehicles Symposium 2000, Dearborn, MI, USA, 5 October 2000; pp. 302–307.
22. Pellkofer, M.; Lützeler, M.; Dickmanns, E.D. Interaction of Perception and Gaze Control in Autonomous Vehicles. In *Intelligent Robots and Computer Vision XX: Algorithms, Techniques, and Active Vision*; Society of Photo Optical: Bellingham, WA, USA, 2001; pp. 1–12.
23. 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. In Proceedings of the Computer Vision Systems, Vancouver, LM, Canada, 7–8 July 2001.
24. Pellkofer, M.; Dickmanns, E.D. Behavior Decision in Autonomous Vehicles. In Proceedings of the Intelligent Vehicle Symposium, 2002. IEEE, Versailles, France, 17–21 June 2002.
25. Damasio, A.R. *Descartes' Irrtum. Fühlen, Denken Und Das Menschliche Gehirn*; Paul List Verlag: Berlin, Germany, 1995; 384p.
26. von Holst, E.; Mittelstaedt, M. Das Reafferenzprinzip. *Naturwissenschaften* **1950**, *37*, 464–476. [CrossRef]
27. Ruhnau, F. The Now—The missing link between matter and mind. In *The Now, Time and Quantum*; Bitbol, M., Ruhnau, E., Eds.; Atlantica Séguier Frontières: Biarritz, France, 1994.
28. Ruhnau, F. The Now—A hidden window to dynamics. In *Inside versus outside. In Endo- and Exo-Concepts of Observation and Knowledge in Physics, Philosophy and Cognitive Science*; Atmanspacher, A., Dalenoort, G.J., Eds.; Springer: Berlin/Heidelberg, Germany, 1994.
29. Singer, W. *Neurobiology of Human Values*; Springer: Berlin/Heidelberg, Germany, 2005; ISBN 978-3-540-26253-4.
30. Singer, W. *Dynamic Coordination in the Brain: From Neurons to Mind*; MIT Press: Cambridge, MA, USA, 2010; ISBN 978-3-642-11667-4.
31. von Heiseler, T.N. Language evolved for storytelling in a super-fast evolution. In Proceedings of the 10th International Conference (EVOLANG 10), London, UK, 20 May 2014; pp. 114–121.
32. Albus, J.S.; Meystel, A.M. *Engineering of Mind—An Introduction to the Science of Intelligent Systems*; J. Wiley & Sons Publication: Hoboken, NJ, USA, 2001; 411p.
33. Albus, J.S. 4-D/RCS reference model architecture for unmanned ground vehicles. In Proceedings of the IEEE International Conference on Robotics and Automation, San Francisco, CA, USA, 24–28 April 2000.
34. Schlenoff, C.; Albus, J.; Messina, E.; Barbera, A.J.; Madhavan, R.; Balakirsky, S. Using 4D/RCS to Address AI Knowledge Integration. *Ai Mag.* **2006**, *27*, 71.
35. Kröger, F. *From Automated to Autonomous Driving. A Transnational Research History on Pioneers, Artifacts and Technological Change (1950–2000)*; Springer: Berlin/Heidelberg, Germany, 2024.
36. Parekh, D.; Poddar, N.; Rajpurkar, A.; Chahal, M.; Kumar, N.; Joshi, G.P.; Cho, W. A Review on Autonomous Vehicles: Progress, Methods and Challenges. *Electronics* **2022**, *11*, 2162. [CrossRef]
37. Roland, A.; Shiman, P. *Strategic Computing: DARPA and the Quest for Machine Intelligence, 1983–1993*; MIT Press: Cambridge, MA, USA, 2002.

38. Basic Research: PRO-ART: Methods and Systems of Artificial Intelligence; CED 3: Collision Avoidance. Available online: https://en.wikipedia.org/wiki/Eureka_Prometheus_Project (accessed on 14 October 2024).
39. Leontyev, A.N. *The Development of Mind—Selected Works of Aleksei Nikolaevich Leontyev*; Bookmasters, Inc.: Ashland, OH, USA, 2009.
40. Dickmanns, E.D. Developing the Sense of Vision for Autonomous Road Vehicles at the UniBwM. *Computer* **2017**, *50*, 24–31. [CrossRef]
41. Pomerleau, D.A. Neural Netwok Perception for Mobile Robot Guidance. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA, February 1992.
42. Kocić, J.; Jovičić, N.; Drndarević, V. An End-to-End Deep Neural Network for Autonomous Driving Designed for Embedded Automotive Platforms. *Sensors* **2019**, *19*, 2064. [CrossRef] [PubMed]
43. Hussain, R.; Zeadally, S. Autonomous Cars: Research Results, Issues, and Future Challenges. *IEEE Commun. Surv. Tutor.* **2018**, *21*, 1275–1313. [CrossRef]
44. Bachute, M.R.; Subhedar, J.M. Autonomous Driving Architectures: Insights of Machine Learning and Deep Learning Algorithms. *Mach. Learn. Appl.* **2021**, *6*, 100164. [CrossRef]
45. Li, Z.; Du, Y.; Zhu, M.; Zhou, S.; Zhang, L. A survey of 3D object detection algorithms for intelligent vehicles development. *Artif. Life Robot.* **2021**, *27*, 115–122. [CrossRef] [PubMed]
46. Lee, D.J.; Zhang, D. Special Issue Editors: Convolutional Neural Networks and Vision Applications, Volume II. 2022.
47. Yurtsever, E.; Lambert, J.; Carballo, A.; Takeda, K. A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE Access* **2020**, *8*, 58443–58469. [CrossRef]
48. Arikumar, K.S.; Deepak Kumar, A.; Gadekallu, T.R.; Prathiba, S.B.; Tamilarasi, K. Real-Time 3D Object Detection and Classification in Autonomous Driving Environment Using 3D LiDAR and Camera Sensors. *Electronics* **2022**, *11*, 4203. [CrossRef]
49. Liu, Y.; Hu, H.; Chen, D. Attentive Part-Based Alignment Network for Vehicle Re-Identification. *Electronics* **2022**, *11*, 1617. [CrossRef]
50. Pihrt, J.; Simánek, P. Spatiotemporal Prediction of Vehicle Movement Using Artificial Neural Networks. In Proceedings of the 2022 IEEE Intelligent Vehicles Symposium, Aachen, Germany, 4–9 June 2022; pp. 734–739.
51. Huch, S.; Sauerbeck, F.; Betz, J. DeepSTEP—Deep Learning-Based Spatio-Temporal End-To-End Perception for Autonomous Vehicles. In Proceedings of the 2023 IEEE Intelligent Vehicles Symposium, Anchorage, AK, USA, 4–7 June 2023.
52. Liu, C.; Li, X.; Liu, Q.; Li, Z.; Li, M. A Review of Vision-Based Road Detection Technology for Unmanned Vehicles. In Proceedings of the 2023 IEEE Intelligent Vehicles Symposium, Anchorage, AK, USA, 4–7 June 2023.
53. Autonomous Driving in Traffic with End-to-End Vision-Based Deep Learning 2024. Available online: https://www.sciencedirect.com/science/article/pii/S0925231224006453 (accessed on 14 October 2024).
54. Ebert, F.; Wuensche, H.-J. Dynamic Object Tracking and 3D Surface Estimation using Gaussian Processes and Extended Kalman Filter. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference, Auckland, New Zealand, 27–30 October 2019.
55. Forkel, B.; Kallwies, J.; Wuensche, H.-J. Combined Road Tracking for Paved Roads and Dirt Roads: Framework and Image Measurements. In Proceedings of the 2021 IEEE Intelligent Vehicles Symposium, Nagoya, Japan, 11–17 July 2021.
56. Forkel, B.; Wuensche, H.-J. Dynamic Resolution Terrain Estimation for Autonomous (Dirt) Road Driving Fusing LiDAR and Vision. In Proceedings of the 2022 IEEE Intelligent Vehicles Symposium, Aachen, Germany, 4–9 June 2022.
57. Forkel, B.; Wuensche, H.-J. Combined Road Tracking for Paved Roads and Dirt Roads: LiDAR Measurements and Image Color Modes. In Proceedings of the 25th International Conference on Information Fusion (FUSION), Linköping, Sweden, 4–7 July 2022.
58. Bienemann, A.; Wuensche, H.-J. Model Predictive Control for Autonomous Vehicle Following. In Proceedings of the 2023 IEEE Intelligent Vehicles Symposium, Anchorage, AK, USA, 4–7 June 2023.
59. Friston, K. The free-energy princip. In *Open MIND*; Metzinger, T.K., Windt, J.M., Eds.; MIND Group: Frankfurt am Main, Germany, 2010. [CrossRef]
60. Menary, R. Cognitive integration, enculturated cognition and the socially extended mind. *Cogn. Syst. Res.* **2013**, *25–26*, 26–34. [CrossRef]
61. Hohwy, J. *The Predictive Mind*; Oxford University Press: Oxford, UK, 2013.
62. Hohwy, J. The self-evidencing brain. *Noûs* **2016**, *50*, 259–285. [CrossRef]
63. Fabry, R.E. Predictive processing and cognitive development. In *Philosophy and Predictive Processing*; Metzinger, T., Wiese, W., Eds.; Bloomsbury Publishing Inc.: New York, NY, USA, 2017; Volume 11, pp. 127–138.