Knowledge Bases for Visual Dynamic Scene Understanding

Ernst D. Dickmanns
Technology of Autonomous Systems, University of the Bundeswehr, Munich, Germany
Department of Aero-Space Technology (LRT), Neubiberg, Germany

Keywords: Knowledge Representation, Real-time Machine Vision, Behavior Decision, Scene Understanding.

Abstract: In conventional computer vision the actual 3-D state of objects is of primary interest; it is embedded in a temporal sequence analyzed in consecutive pairs. In contrast, in the 4-D approach to machine vision the primary interest is in temporal processes with objects and subjects (defined as objects with the capability of sensing and acting). All perception of 4-D processes is achieved through feedback of prediction errors according to spatiotemporal dynamical models constraining evolution over time. Early jumps to object/subject-hypotheses including capabilities of acting embed the challenge of dynamic scene understanding into a richer environment, especially when competing alternatives are pursued in parallel from beginning. Typical action sequences (maneuvers) form an essential part of the knowledge base of subjects. Expectation-based Multi-focal Saccadic (EMS-) vision has been developed in the late 1990s to demonstrate the advantages and flexibility of this approach. Based on this experience, the paper advocates knowledge elements integrating action processes of subjects as general elements for perception and control of temporal changes, dubbed 'maneuvers' here. − As recently discussed in philosophy, emphasizing individual subjects and temporal processes may avoid the separation into a material and a mental world; EMS-vision quite naturally leads to such a monistic view.

1 INTRODUCTION

Starting in ancient philosophy (Socrates, Plato, Aristotle), then continued every now and then over almost 2500 years, and especially in the recent past there has been a discussion about what is the right way to treat the phenomenon of 'knowledge'. Is there a quasi-static truth over and above individuals (Plato and followers) or do we essentially observe physical processes and actions of individuals and then have to come to interpretations that have to be mutually accepted without any guaranteed relation to an observer-independent truth.

In their dissertations (Noe 1995, see Noe 2004; Kiverstein 2005) the authors investigate the question whether the claimed insurmountable gap between the naturalistic and the idealistic (phenomenal) philosophical view can be bridged by an approach basically different from the predominant one. They came to the conclusion that avoiding the quasi-static 'absolute' view and relying more on temporal processes with individual subjects may eliminate the development of a gap. In a similar direction hint the results of the Russian psychologist and philosopher (Leontyev 2009). In his view the core of knowledge is the capability of individuals to make sense of a process observed and to respond with some activity, the outcome of which is the basis for learning behaviors and for developing capabilities for goal oriented decisions as well as for a system of values.

In terms of modern neurophysiology this view may be associated with the effect of the *mirror neurons* in brains of vertebrates that are active both when an action is performed and when it is visually observed by a subject (Gallese and Goldman 1998, see also *web-entries*). The essential point is that activities over time are directly represented in neural systems as well as abstracted quasi-static results.

In the field of 'Cognitive Vision', a survey may be found in (Christensen and Nagel 2006) with over 500 references. The introductory Section (Vernon 2006) finishes with the conclusion: "Broadly speaking, there are essentially two approaches to recognition: 1. The cognitivist symbolic information processing representational approach, 2. The emerging systems approach (connectionism, dynamical systems, enactive systems)... . The former one takes a predominantly static view of knowledge

D. Dickmanns E..
Knowledge Bases for Visual Dynamic Scene Understanding.
DOI: 10.5220/0005340802090215
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 209-215
ISBN: 978-989-758-089-5

represented by symbol systems that refer to the physical reality that is external to the cognitive agent. ... The emergent systems approach: ● takes a predominantly dynamic or process view of knowledge, and views it more as a collection of abilities that encapsulate 'how to do' things; ● is therefore subservient to the cognitive agent and dependent on the agent and the environmental context."

The former (quasi-static) approach was the predominant one in the 1980s. A group at UniBw Munich has used a dynamical-systems-approach for vehicle guidance by computer vision by relying on feedback of prediction errors for image features (Kalman 1960). The use of real-world dynamical (4-D) models for image sequence processing resulted in a breakthrough in performance achieved with very limited computing power (Dickmanns, Graefe 1988, Dickmanns 2007). For road vehicle guidance by machine vision, in the meantime, the approach based on the *Extended Kalman Filter* (EKF) has become the standard method for recognition of road and lane parameters as well as for tracking other vehicles.

In any case it is essential that the dynamical models used represent objects in the real world, including their behavior over time (and not in some intermediate measurement space like image coordinates). Human knowledge about the world is mainly geared to objects and classes of object. Two hyper-classes of objects have to be distinguished if scene understanding on the semantic level is the goal: 1. Objects that are not able to initiate motion on their own (called here more precisely: 'objects proper') and 2. objects that are able to sense information about the environment and to activate some control output affecting their physical state; the latter will be dubbed **'*subjects*'** here. All animals and robots fall into this category. The simple term 'object' is used here for both types.

For understanding of scenes including subjects it is mandatory to have knowledge available about how these subjects transform their sensor data into own behavior. If this triggering of behavior is not a fix program, like in humans, the closed-loop sequence of sensing, behavior decision and acting is of importance. Since direct access to mental processes of subjects is not possible, the best substitute is to try to grasp a subject's intention by observing the onset of maneuvers. This is possible only if typical maneuvers of members of the class of subjects observed are represented in the knowledge base of the observer; in the context of the situation given, the likely candidates for maneuvers have to be recognized from data of their onset. Knowledge is not considered to be absolute truth but the best background available in the individuals or in the community for arriving at proper decisions in actual or future situations of any kind.

2 EFFICIENT REPRESENTATION OF SUBJECTS

To a large extent, knowledge about the world is linked to classes of subjects and to their individuals. Beside geometrical shape and body articulation these classes and their individuals are characterized by the capabilities of: a) sensing, b) data processing and perception on a higher mental level, c) decision making in a situational context, and d) control actuation for achieving some goal or a mission.

In order to understand the semantics of what these individuals are doing it is necessary to have knowledge about the maneuvers performed and about the context these maneuvers are applied in, usually. This means that three levels should be represented and used in parallel:

1. The **visual feature level** with links to objects and to how their motion and ego-motion affect the appearance of these features (Jacobian matrices);
2. The **object level** with:
   α) Body shape and articulation,
   β) typical movements of limbs, head/neck and the body as part of maneuver elements for locomotion or some other goal.
   γ) Feature distribution on their 3-D surface.
   δ) Typical goals of subjects in given situations.
3. The task domain on the **situation level** with typical environmental conditions.

One basic task of subjects is to come up with well-suited decisions for their own behavior given the environmental conditions perceived and the own system of goals and values. Thus, the whole range from features of objects to **situations for subjects** has to be considered in parallel.

2.1 The Decision Framework

Figure 1 visualizes the ranges needed both in 3-D space (vertical in first column, range elements in blue) and in 1-D time (horizontal in first row, range elements in red). All measurements are done at the point 'here and now' in the upper left corner of the yellow rectangle. Since the sensors are distributed over the vehicle, usually, the effect of the dislocation of each sensor from the center of gravity (cg) as the point of reference for motion has to be taken into account. The worst effect due to dislocation is

experienced in inertial sensing with accelerometers. Their signals contain beside the acceleration at the cg also components from rotational accelerations and from centrifugal forces due to rotational speeds. These signals, however, are available at almost no time delay (microseconds range) and contain the effects of any type of perturbations on the vehicle body directly, like hitting a pothole or wind gusts.

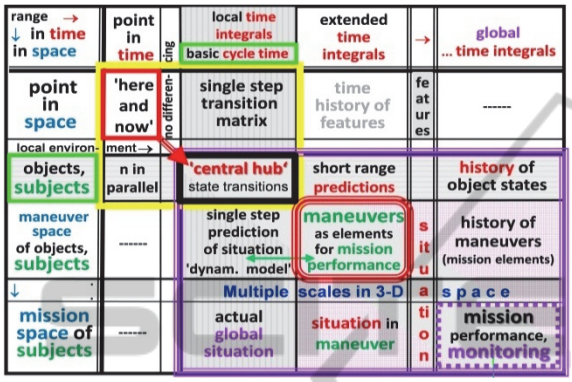

Figure 1: Multiple scales in space (vertical) and time (horizontal) for recognition and tracking of image features, objects / subjects, and the situation in real-time image sequences including results from inertial sensors.

On the contrary, video signals allow discovering these perturbations only from careful observation of integrals of these effects that are linked to temporal derivatives of vehicle states (positions and angular orientations); the analysis of image sequences includes time delays of several video cycles (~ 0.1 to 0.3 sec.). In addition, high angular rates may lead to motion blur in images. This is the reason why advanced biological vision systems like those of vertebrates have combined visual / inertial sub-systems that allow exploiting the advantages and avoiding the disadvantages of both systems used separately. Negative feedback of inertial angular rate data, sensed on the base of the gaze platform, onto the commanded gaze direction reduces the amplitudes of rotational perturbations in the images by more than an order of magnitude, thereby alleviating image interpretation (Dickmanns 2015). On the other hand, the tendency towards drift errors resulting from continuous integration of inertial rate data can easily be counteracted by visual feedback based on proper edge or corner features from stationary objects far away. These properties of combined interpretation of inertial and visual sensor data may have given rise to developing a feeling (and later on the notion) of time and temporal integrals as essential elements of knowledge in dynamic scene understanding of biological systems. This aspect has been neglected in many approaches to real-time machine vision for motion control. Proper handling of delay times and corresponding treatment of the effects of time integrals using dynamical models has been an important ingredient to the early successes of the 4-D approach to real-time machine vision with low computing power.

Taking conventional measurement signals for vehicle speed, distance traveled, and steer angle into account in the framework of full dynamical models even allows *monocular motion stereo* interpretation at almost no extra cost. This is equivalent to the highest level of 'Self-Localization And Mapping' (SLAM) and even more exact than the so called '6D approach' in this field (where the D means 'degree of freedom' and not 'dimension' as in the 4-D approach). The central hub shown in the center of Figure 1, where the 'object'-row and the center-column representing 'basic video cycle time' intersect each other, combines all actual information on objects perceived including the full state of the own subject. Since in recursive estimation by feedback of prediction errors of features temporal differentiation of noisy sensor data is avoided (horizontal center of the yellow rectangle) but a smoothing integration step is used, this approach is superior to inverse perspective projection based on two consecutive images.

The standard nonlinear equations of perspective projection are linearized around the actual state. The so called '*Jacobian matrix*' then linking parameters of visual features linearly to object state components has to be inverted for obtaining better state estimates from prediction errors of features. This matrix is an important knowledge element since it contains the information how a feature will change in the image if a state or shape component of the object is varied. This is the reason why Extended Kalman Filtering (EKF) for vision is preferred over particle filtering or other variants if the objects in the scene are known to sufficient detail; this check has to be done on the level of task domains and of potential situations encompassing a large number of different objects and environmental conditions. Since this involves large ranges in both space and time, it is shown in Figure 1 in the lower right corner (rectangle in magenta); here, only abstracted data from n objects tracked in parallel and abstracted situational data are of importance. By making the transition from image data to objects, the volume of data is reduced by two to three orders of magnitude, hopefully without losing relevant information on essential components in the scene. This allows checking situations with many individual objects by

referring to proper knowledge bases containing the behavioral capabilities and possibly the preferences of subjects in certain conditions. Three levels of knowledge bases linked to the main diagonal of Figure 1 will be discussed below.

2.2 Visual Features

At the point 'here & now' primary feature detection can be done purely bottom-up without reference to previous images; only local neighborhoods in the image plane are taken into account. This yields features like: 1. local regions with nonplanar intensity distributions, shown in Figure 2 in white; 2. edge elements, shown in red (vert. search) and green (hor. search), 3. Corners (blue crosses), and 4. larger regions with homogeneous gray shading, (colors or textures have not been evaluated in Figure 2).

Many types of additional features derived from object hypotheses may be used during tracking phases in a feedback mode of prediction-errors using recursive estimation methods. Typical examples are to look for wheels (usually parts of ellipses) and tires (dark) or for groups of head- and backlights relative to the position of vehicle bodies.

2.3 Objects / Subjects in Motion

A human observer looking at the synthetic image in Figure 2 cannot but immediately recognize a three-lane road with heavy traffic. The gray regions in the lower part ('nearby') with typical 'lane markings' (both the white local regions as locations of non-planar gray value distribution and the red-colored edge elements within them, forming almost-straight longer line segments) enforce this interpretation. For an experienced human driver seven objects above the road are readily detected. Usually it takes three to seven video cycles (~ 0.1 to ~ 0.3 seconds) to achieve a stable interpretation with small sums of squared prediction errors in road scenes.

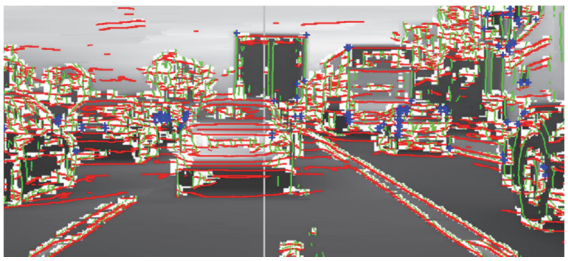

Figure 2: Scene with dense traffic made up of five types of features (no single pixels!).

The additional degrees of freedom of subjects relative to 'objects proper' require that for scene understanding the latter and 'subjects' have to be treated differently. While for 'objects proper' knowledge about laws of motion is sufficient, for subjects the self-decided variation of movements is an additional degree of complexity for adequate perception and understanding of motion processes.

The distinction for dynamical systems between *state variables* and *control variables*, introduced by Caratheodory in the first half of the last century for treating optimal control problems, may be the key to better understanding of movements of subjects. The following definition holds: *State variables* in a dynamical system are all those variables, the value of which cannot be changed at one point in time; *they evolve over time* with differential equations describing the constraints holding. If formulated properly, the state variables contain all effects of the past. Only the actual state and the control variables – to be chosen freely (within limits) at each moment – determine its future development. It is interesting to note that the presence of control variables in dynamical systems is a precondition for developing a free will. If there is no control variable available in a system, its future development cannot be influenced (and thus 'free will' is meaningless).

On the other side, if there are no measurement data available, there is no base for proper decision making and application of behaviors in the (then unknown) environment given. Thus, the 'sensing – acting loop closures' are the driving factors for '*mental data processing of subjects*'. Beside body shape and articulation as well as kind of locomotion it is the capability of sensing and data processing that determines the class of subjects among animals.

A rather direct link from sensed data to control actuation is dubbed a reflex. The idea of evolution is that during this process more and more senses have developed providing various data in parallel. Those animals that happened to use them in a fashion leading to superior results for their survival in the environment encountered, had better chances to generate more descendants and to spread. Combined use of data from separate paths must have been one important step of development. This process has generated a multitude of classes of living beings. Most of them developed specialized organs for the combination of sensor data and finally the brain.

Visual perception of the environment plays a dominant role in the development of cognitive capabilities. In the neural systems of vertebrates, data processing and cognition is based on temporal processes in billions of neurons with very many cross-connections. Detailed functioning of this very

complex biochemical / electrical network is yet widely unknown. However, it seems likely that frequently observed *typical motion processes* of other objects or subjects form part of the knowledge base for understanding of situations. In biological systems these *maneuvers* are learned by repeated observation or by own exercises.

Especially in the latter case it is not the trajectory of the body and the limbs that are learned but the *time history of the control output* leading to these trajectories. This procedure is a much more efficient encoding of the maneuver for application since it concentrates on those variables that are the only ones to be changed directly. Guiding a road vehicle for a lane change thus does not require a trajectory to be stored (with ~ half a dozen state variables over extended ranges in time) but just the (parameterized) time history of the one control variable "steer angle rate" to be applied. Properly scaled, in the nominal case without perturbations, this needs less than a dozen numbers for the entire maneuver.

Since in the real world perturbations both in the environment (cross-winds, road sloping, pot holes, etc.) and in the perception system are more the rule than an exception, superimposed feedback control for counteracting these effects is mandatory. The reference trajectory for this feedback component can be computed online from the nominal parameters actually used in the dynamical model of the maneuver. With little additional computing effort this also provides the coefficients for linear state feedback control that yields acceptable eigenvalues for the closed-loop system due to the knowledge stored in the dynamical model (Dickmanns 2007).

2.4 Situations in Task Domains

A 'situation' is defined as the complete collection of all conditions relevant for decision making for a subject. It encompasses all relevant environmental conditions in the task domain: Weather conditions, lighting- and visibility conditions, surface conditions for ground vehicles, local geometrical structure and objects around. In all cases the mission to be performed and its decomposition into a list of consecutive mission elements and maneuvers are stored symbolically; timing conditions for transitions and the own health state are of importance. All potential situations constitute such a tremendous volume that subdivision into specific task domains is mandatory. In human society, this and the limited capabilities of single individuals are the reason for the development of the many existing professions.

Within each task domain there are characteristic missions to be performed; each mission can be subdivided into a sequence of mission elements that can be treated with the same set of behavioral components. Certain *maneuvers* are characteristic for specific mission elements and for the transition between those; their *proper representation* is essential for efficient overall systems.

There is also a need for evaluating the performance levels achieved and for keeping track of their changes over time under different environmental conditions (both improvements and deteriorations). These values form the basis for adapting maneuver parameters and for selecting maneuvers in the future in accordance with the situation encountered. This constitutes learning of (dynamical) behavioral components; learning which one of these parameter sets should be used in which situations is what constitutes 'experience in the field'. This experience allows recognizing snapshots as part of a process; on this basis expectations can be derived that allow a) focusing attention in feature extraction on special events (like occlusion or uncovering of features in certain regions of future images) or b) increased resolution in some region of the real world by gaze control for a multifocal system (dashed curves in lower left of Figure 3).

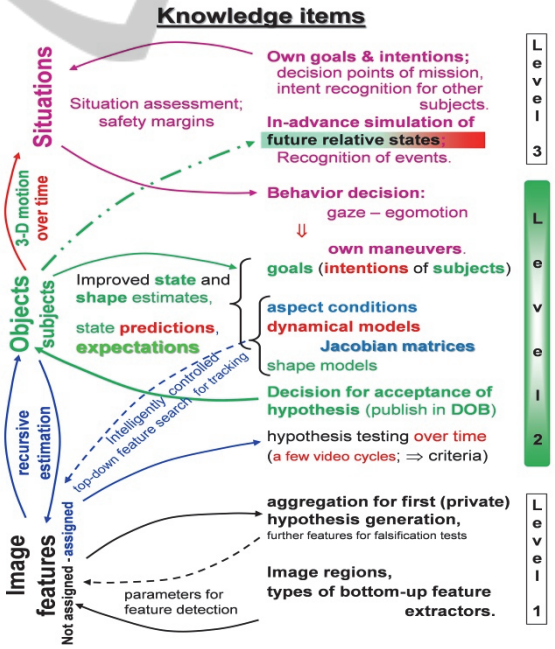

Figure 3: Information flow between the three levels of knowledge representation in dynamic vision.

Crucial situation-dependent decisions have to be made for transitions between mission phases where switching between behavioral capabilities for

maneuvers is required. That is why representation of specific knowledge for '*maneuvers*' is important (rounded central red block within the rectangle in the lower right corner of Figure 1).

3 THREE LEVELS IN PARALLEL

As mentioned previously, the levels discussed separately above have to be treated in parallel with continuous feedback between them. Figure 3 sketches the information flow. At the base are consecutive image evaluation processes *independent of temporal aspects*. However, a component for the generation of object hypotheses has to be available interpreting collections of features that might stem from the same real-world object. Initially, this hypothesis is kept locally private for testing over the next few video cycles. Only after the sum of the squared prediction errors remains below a threshold, the hypothesis is made public in the perception system by inserting it into the scene tree representing the relative states by homogeneous coordinates. This makes the objects available to situation-level 3. [For more detailed discussions see (IV'00, 2000), Chap. 13 of (Dickmanns 2007), and www.dyna-vision.de].

With the object states given in the scene tree and with the actions of subjects assumed to be performed, a single-step prediction of the states for the next point in time of measurements is computed (text in red in Figure 3). This allows intelligent control of top-down feature search (dashed arrow in blue). For objects of special interest, longer range predictions may be made for extended situation analysis (green dash-dotted arrow) to the top level 3. There may be separate routines for perceiving and representing environmental conditions that may need evaluation of special features (like decreasing contrast with visual range under foggy conditions).

At the situation level (top in Figure 3), all of this information is evaluated in conjunction, and the result is communicated to the two sublevels for control of gaze direction and own locomotion in the mission context.

4 SUMMARY OF POSITION

Experience in joint use of procedural methods from 'Control Engineering' and declarative methods from 'Artificial Intelligence' for processing of image sequences and for scene understanding has led to the proposal to expand the knowledge base for dynamic real-time vision and control of actions by a specific component for '*maneuvers*': Such a component for the transition from state $S_1(t_1)$ to $S_2(t_2)$ contains for each of these mission element (S_1 to S_2) in task domains the following information:

- The nominal control time histories $\underline{\mathbf{u}}(\cdot)$;
- the dynamical model for generating the nominal trajectories of the state variables;
- code for generating the coefficients of feedback control laws for counteracting perturbations,
- conditions under which the *maneuver* may be used with which set of parameters.
- Codes for evaluating pay-off functions that allow judging the quality of the maneuver performed.

This process-oriented approach geared to the control variables of dynamical systems is more efficient than centering on state variables.

REFERENCES

Christensen H. I., Nagel H.-H. (eds.), 2006. Cognitive Vision Systems – Sampling the Spectrum of Approaches. *Springer,* (367 papes).

Dickmanns, E.D., 2007. Dynamic Vision for Perception and Control of Motion. *Springer* (474 pages).

Dickmanns, E.D., 2015. BarvEye: Bifocal active gaze control for autonomous driving. (this volume).

" , Graefe, V., 1988. a) Dynamic monocular machine vision. *Machine Vision and Applications, Springer International, Vol. 1*, pp 223-240. b) Applications of dynamic monocular machine vision. pp 241-261.

Gallese V., Goldman A. 1998. Mirror Neurons and the Simulation Theory of Mind-reading. *Trends in Cogn. Sci.2*, pp 493-501.

IV'00, 2000. Proc. Internat. Symp. on Intelligent Vehicles, Dearborn (MI), with six contributions to Expectation-based, Multi-focal, Saccadic (EMS-) vision:

1. Gregor R. et al.: EMS-Vision: A Perceptual System for Autonomous Vehicles.
2. Gregor R., Dickmanns E.D.: EMS-Vision: Mission Performance on Road Networks.
3. Hofmann U.; Rieder A., Dickmanns, E.D.: EMS-Vision: Applic. to 'Hybrid Adaptive Cruise Control'.
4. Luetzeler M., Dickmanns E.D.: EMS-Vision: Recognition of Intersections on Unmarked Road Networks.
5. Pellkofer M., Dickmanns E.D.: EMS-Vision: Gaze Control in Autonomous Vehicles.
6. Siedersberger K.-H., Dickmanns E.D.: EMS-Vision: Enhanced Abilities for Locomotion.

Kalman, R. D. 1960. A new approach to linear filtering and prediction problems. *Trans. ASME, Series D, Journal of Basic Engineering,* pp 35–45.

Kiverstein J.D., 2005. Naturalism and Phenomenology. *Diss. Univ. Edinborough.*

Leontyev A. N. 2009. The Development of Mind.

(Selected Works). *Marxists Internet Archive. Printed by Bookmasters, Inc., Ohio.*

Noe Alva 2004. Action in Perception. *Cambridge, MA; MIT Press.*

Vernon D., 2006. The Space of Cognitice Vision. *In Christensen and Nagel (eds.)*, pp 7-24.