Dickmanns ED: An **E**xpectation-based, **M**ulti-focal, **S**accadic (EMS) (inertially stabilized dynamic) **Vision System for Vehicle Guidance**

Extended abstract: After a brief review of elder system architectures for bifocal vision and their deficiencies, the 'MarVEye' camera configuration realized with three to four conventional miniature CCD-TV-cameras mounted fix relative to each other on a high bandwidth pan-and-tilt platform will be discussed. Two wide-angle cameras, with divergent optical axes in a plane arranged such that there is a central region of overlap, provide both a wide simultaneous horizontal field of view (> ~100°) and a central area for binocular stereo evaluation. A third camera with a mild tele-lens covering the central part of the stereo region in vertical direction even allows robust trinocular stereo interpretation in this field of view; it has been chosen as a 3-chip-color-camera with an aperture of about 15°. The optical axes of these three cameras lie in a plane. Only edge feature extraction is done at present for fast and easy correction of distortions in the wide-angle images (after feature localization to subpixel accuracy) for stereo interpretation.
A fourth camera with an even larger focal length may be added for good performance in recognition further away (f = ~75 mm), allowing a resolution of about 5 cm per pixel at a distance of 400 meter. This black-and-white camera may have a high sensitivity with respect to light intensity in order to obtain complementary photometric properties in the camera set.
Exploitation of the capabilities thus given requires intelligent viewing direction control. Systematic search patterns are available for object detection. Visually locking onto single moving objects is another mode for reduction of motion blur and for better performance in recognition of details; the area thus covered is relatively small, however. If other objects are being noticed in the wide field of view, a fast saccade may be made in order to get one of these objects into better focus. Only a few vision cycles (of 40 ms duration) may be sufficient for achieving object classification; then, further tracking of this object may again be done in the wide-angle images. This saccadic type of vision with intermittent phases of fast saccades and smooth pursuit does complicate image sequence interpretation. However, with explicit spatio-temporal representations available in the 4-D approach this can be handled in a straightforward way.

Joint inertial and visual perception with mutual crosswise stabilization has been realized for this purpose. Inexpensive angular rate sensors on the moving part of the platform for viewing direction control, accurately pick up rotations in a frequency band ranging up to several Hz; the high frequency part of the signal is sufficiently good, while low frequency drifts need special handling. Commanding the negative value of the rate actually measured to the viewing direction control leads to stabilization of the viewing direction with an accuracy of a few tenths of a degree. On a pitching vehicle, this means that a large part of the motion blur induced by this pitching will be removed, thereby alleviating image processing. In land vehicles, where the viewing direction is almost parallel to the ground, this also means that the lines of the stabilized images correspond to almost constant look-ahead distances. Slow inertial drifts can easily be stabilized by visual feedback from stationary objects far away (e.g. horizon or some landmark). Constant inertial scan rates can thus also be realized despite perturbations.
A set of three mutually orthogonal accelerometers allow picking up all accelerations on the own body which will yield linear velocities (first integrals over time) and positions (second integral). Perspective mapping of other objects into the image plane depends on these relative positions (second integrals). Thus, inertial sensing yields lead information for image processing including the effects of unforseeable perturbations. By taking all six (translational and rotational) degrees of freedom into account, short-term predictions can be made without any analytical models for the dynamics of the rigid body, except the relationships of integration over time. Note, that temporal models are unavoidable for this type of data fusion; the 4-D approach makes use of them right from the beginning. In biological systems like vertebrates, this interaction of inertial

and visual sensing is well known from several data paths between vestibular and ocular subsystems.
The core representation scheme in the EMS vision system is a stabilized animated world model in 3-D space and time driven by ***inertial and visual*** data feedback. Its central hub is a generic knowledge representation for objects and subjects, the latter ones being defined as objects with the capability of collecting information and generating control actions on their own decision. Differential and integral representations on different scales are used simultaneously.

Multiple scale representations in 3-D space and time being used span several orders of magnitude. In space, it ranges from pixel size (a few micrometers) to global missions on Earth (~ 10 000 km) [even planetary lighting conditions by the Sun, 1.5 x 10^8 km]; in time it ranges from milliseconds for viewing direction control to several hours for mission performance [and years for planetary seasons]. Scaling by a single parameter in homogeneous coordinates (space) or on the time scale is the standard method used. Homogeneous transformations link object positions and orientations in a 'dynamic scene tree'. The trouble is that the entries in the transformation matrices are the unknowns of the vision problem. A systematic approach has been developed to determine these unknowns by recursive estimation making use of linear approximation matrices for the relationships between feature positions in the various images of the camera set and state variables or shape parameters of the objects seen. These Jacobian matrices have to be determined for each object/sensor pair in each cycle. An efficient method has been devised exploiting intermediate results of matrix concatenations for perspective mapping.

Object-oriented programming in C++ has been introduced and nicely fits the generic object classes for the 4-D approach. Subject classes as special cases of object classes with control capabilities at their disposal; they have to be represented with additional capabilities for perception, decision-making and control, all of this possibly on various (time) scales. Knowledge is introduced on three different levels:

1. Intelligent control of feature extraction and hypothesis generation for individual objects from collections of features and their behavior over time.
2. On the object/subject level for generic shape (with parameters for adaptation to the visual data measured), for determining the aspect conditions (relative state in 3-D space), and for the generation of behavioral capabilities (characteristic eigen-motion for objects, stereotypical control modes for subjects). This allows efficient tracking based on prediction error feedback and the realization of the 'Gestalt'-idea in recognition. A reduction in computing power required by orders of magnitude may be achieved by exploiting this knowledge carefully in image sequence processing.
3. On the situation level, multiple objects/subjects are being analyzed in the mission context, taking conjectured behaviors of other subjects into account through fast in-advance-simulations over a few seconds of their most likely alternatives.

These new scene tree representations with dynamic management of objects have lead to much improved flexibility and generality. Own decisions for mission performance are based on local maximization of performance criteria along precomputed mission plans with certain degrees of flexibility for adaptation to the situations actually encountered. A dynamic database forms the isolation layer between the control engineering and the AI-oriented methods jointly used. Perceptual and behavioral capabilities are, usually, represented on both levels in different form according to the way they are used. They occur as quasi-static capabilities to be activated in certain situations on the AI-level (automaton states), and as procedural capabilities for actual implementation taking fast data feedback into account (on the control-output level).
The system has been realized on commercially available off-the-shelf PC-hard- and software (four Dual PentiumPro -/II). System integration takes time delays up to several tenths of a

second (for communication and computation from measurement till actuator output) into account. Delay compensations are computed for different data paths. Temporal modeling in the 4-D approach provides all the background information needed.

The system is being applied to road vehicle guidance, first with the 5-ton van VaMoRs for driving on networks of minor roads, and second with the Mercedes S-class test vehicle VaMP for high speed driving on corresponding roads.
Mission performance in nap-of-the-Earth flying has been done with helicopter landmark navigation. The system uses both GPS, inertial, visual and other conventionally measured information (e.g. from speedometer and radar altimeter) for autonomously performing a small mission around the airport of Brunswick with a final landing approach to the heliport marked by the capital letter 'H' on a taxiway. This system has been tested with a real vision system in a real-time hardware-in-the-loop facility in 1997.
Onboard autonomous landing approaches with an in-flight-simulator (twin-jet ATTAS of DLR) for future military transport aircraft are under preparation.

Dynamic machine vision is going to achieve a new level of performance. Increasing computing power by an order of magnitude every 4 – 5 years will allow real-world applications not too far into the future. The 4-D approach naturally lends itself for realizing in technical systems the equivalent of vertebrate vision in biology.