Dynamic Vision for Perception and Control of Motion

Ernst D. Dickmanns

Dynamic Vision for Perception and Control of Motion

Springer

Ernst D. Dickmanns, Dr.-Ing.
Institut für Systemdynamik und Flugmechanik
Fakultät für Luft- und Raumfahrttechnik
Universität der Bundeswehr München
Werner-Heisenberg-Weg 39
85579 Neubiberg
Germany

British Library Cataloguing in Publication Data
Dickmanns, Ernst Dieter
Dynamic vision for perception and control of motion
1. Computer vision - Industrial applications 2. Optical detectors 3. Motor vehicles - Automatic control 4. Adaptive control systems
I. Title
629'.046
ISBN-13: 9781846286377

Library of Congress Control Number: 2007922344

ISBN 978-1-84628-637-7 e-ISBN 978-1-84628-638-4 Printed on acid-free paper

9 8 7 6 5 4 3 2 1

Springer Science+Business Media
springer.com

Preface

During and after World War II, the principle of feedback control became well understood in biological systems and was applied in many technical disciplines to relieve humans from boring workloads in systems control. N. Wiener considered it universally applicable as a basis for building intelligent systems and called the new discipline "Cybernetics" (the science of systems control) [Wiener 1948]. Following many early successes, these arguments soon were oversold by enthusiastic followers; at that time, many people realized that high-level decision–making could hardly be achieved only on this basis. As a consequence, with the advent of sufficient digital computing power, computer scientists turned to quasi-steady descriptions of abstract knowledge and created the field of "Artificial Intelligence" (AI) [McCarthy 1955; Selfridge 1959; Miller *et al.* 1960; Newell, Simon 1963; Fikes, Nilsson 1971]. With respect to achievements promised and what could be realized, a similar situation developed in the last quarter of the 20th century.

In the context of AI also, the problem of computer vision has been tackled (see, *e.g.*, [Selfridge, Neisser 1960; Rosenfeld, Kak 1976; Marr 1982]. The main paradigm initially was to recover a 3-D object shape and orientation from single images (snapshots) or from a few viewpoints. On the contrary, in aerial or satellite remote sensing, another application of image evaluation, the task was to classify areas on the ground and to detect special objects. For these purposes, snapshot images, taken under carefully controlled conditions, sufficed. "Computer vision" was a proper name for these activities since humans took care of accommodating all side constraints to be observed by the vehicle carrying the cameras.

When technical vision was first applied to vehicle guidance [Nilsson 1969], separate viewing and motion phases with static image evaluation (lasting for minutes on remote stationary computers in the laboratory) had been adopted initially. Even stereo effects with a single camera moving laterally on the vehicle between two shots from the same vehicle position were investigated [Moravec 1983]. In the early 1980s, digital microprocessors became sufficiently small and powerful, so that onboard image evaluation in near real time became possible. DARPA started its program "On strategic computing" in which vision architectures and image sequence interpretation for ground vehicle guidance were to be developed ('Autonomous Land Vehicle' ALV) [Roland, Shiman 2002]. These activities were also subsumed under the title "computer vision", and this term became generally accepted for a broad spectrum of applications. This makes sense, as long as dynamic aspects do not play an important role in sensor signal interpretation.

For autonomous vehicles moving under unconstrained natural conditions at higher speeds on nonflat ground or in turbulent air, it is no longer the computer which "sees" on its own. The entire body motion due to control actuation and to perturbations from the environment has to be analyzed based on information coming from many different types of sensors. Fast reactions to perturbations have to be derived from inertial measurements of accelerations and the onset of rotational rates, since vision has a rather long delay time (a few tenths of a second) until the enormous amounts of data in the image stream have been digested and interpreted sufficiently well. This is a well-proven concept in biological systems also operating under similar conditions, such as the vestibular apparatus of vertebrates with many cross-connections to ocular control.

This object-oriented sensor fusion task, quite naturally, introduces the notion of an extended presence since data from different times (and from different sensors) have to be interpreted in conjunction, taking additional delay times for control application into account. Under these conditions, it does no longer make sense to talk about "computer vision". It is the overall vehicle with an integrated sensor and control system, which achieves a new level of performance and becomes able "to see", also during dynamic maneuvering. The computer is the hardware substrate used for data and knowledge processing.

In this book, an introduction is given to an integrated approach to dynamic visual perception in which all these aspects are taken into account right from the beginning. It is based on two decades of experience of the author and his team at UniBw Munich with several autonomous vehicles on the ground (both indoors and especially outdoors) and in the air. The book deviates from usual texts on computer vision in that an integration of methods from "control engineering/systems dynamics" and "artificial intelligence" is given. Outstanding real-world performance has been demonstrated over two decades. Some samples may be found in the accompanying DVD. Publications on the methods developed have been distributed over many contributions to conferences and journals as well as in Ph.D. dissertations (marked "Diss." in the references). This book is the first survey touching all aspects in sufficient detail for understanding the reasons for successes achieved with real-world systems.

With gratitude, I acknowledge the contributions of the Ph.D. students S. Baten, R. Behringer, C. Brüdigam, S. Fürst, R. Gregor, C. Hock, U. Hofmann, W. Kinzel, M. Lützeler, M. Maurer, H.-G. Meissner, N. Mueller, B. Mysliwetz, M. Pellkofer, A. Rieder, J. Schick, K.-H. Siedersberger, J. Schiehlen, M. Schmid, F. Thomanek, V. von Holt, S. Werner, H.-J. Wünsche, and A. Zapp as well as those of my colleague V. Graefe and his Ph.D. students. When there were no fitting multi-microprocessor systems on the market in the 1980s, they realized the window-oriented concept developed for dynamic vision, and together we have been able to compete with "Strategic Computing". I thank my son Dirk for generalizing and porting the solution for efficient edge feature extraction in "Occam" to "Transputers" in the 1990s, and for his essential contributions to the general framework of the third-generation system *EMS* vision. The general support of our work in "control theory and application" by K.-D. Otto over three decades is appreciated as well as the infrastructure provided at the institute ISF by Madeleine Gabler.

Ernst D. Dickmanns

Acknowledgments

Support of the underlying research by the Deutsche Forschungs-Gemeinschaft (DFG), by the German Federal Ministry of Research and Technology (BMFT), by the German Federal Ministry of Defense (BMVg), by the Research branch of the European Union, and by the industrial firms Daimler-Benz AG (now DaimlerChrysler), Dornier GmbH (now EADS Friedrichshafen), and VDO (Frankfurt, now part of Siemens Automotive) through funding is appreciated.

Through the German Federal Ministry of Defense, of which UniBw Munich is a part, cooperation in the European and the Trans-Atlantic framework has been supported; the project "AutoNav" as part of an American-German Memorandum of Understanding has contributed to developing "expectation-based, multifocal, saccadic" (EMS) vision by fruitful exchanges of methods and hardware with the National Institute of Standards and Technology (NIST), Gaithersburgh, and with Sarnoff Research of SRI, Princeton.

The experimental platforms have been developed and maintained over several generations of electronic hardware by Ingenieurbüro Zinkl (VaMoRs), Daimler-Benz AG (VaMP), and by the staff of our electromechanical shop, especially J. Hollmayer, E. Oestereicher, and T. Hildebrandt. The first-generation vision systems have been provided by the Institut für Messtechnik of UniBwM/LRT. Smooth operation of the general PC-infrastructure is owed to H. Lex of the Institut für Systemdynamik und Flugmechanik (UniBwM /LRT/ ISF).

Contents

1 Introduction

The field of “vision” is so diverse and there are so many different approaches to the widespread realms of application that it seems reasonable first to inspect it and to specify the area to which the book intends to contribute. Many approaches to machine vision have started with the paradigm that easy things should be tackled first, like single snapshot image interpretation in unlimited time; an extension to more complex applications may later on build on the experience gained. Our approach on the contrary was to separate the field of dynamic vision from its (quasi-) static counterpart right from the beginning and to derive adequate methods for this specific domain. To prepare the ground for success, sufficiently capable methods and knowledge representations have to be introduced from the beginning.

1.1 Different Types of Vision Tasks and Systems

Figure 1.1 shows juxtapositions of several vision tasks occurring in everyday life. For humans, snapshot interpretation seems easy, in general, when the domain is well known in which the image has been taken. We tend to imagine the temporal context and the time when the image has been shot. From motion smear and unusual poses, the embedding of the snapshot in a well-known maneuver is concluded. So in general, even single images require background knowledge on motion processes in space for more in-depth understanding; this is often overlooked in machine or computer vision. The approach discussed in this book (bold italic letters in Figure 1.1) takes motion processes in “3-D space and time” as basic knowledge required for understanding image sequences in an approach similar to our own way of image interpretation. This yields a natural framework for using language and terms in the common sense.

Another big difference in methods and approaches required stems from the fact that the camera yielding the video stream is either stationary or moving itself. If moving, linear or/and rotational motion also may require special treatment. Surveillance is done, usually, from a stationary position while the camera may pan (rotation around a vertical axis, often also called yaw) and tilt (rotation around the horizontal axis, also called pitch) to increase its total field of view. In this case, motion is introduced purposely and is well controlled, so that it can be taken into account during image evaluation. If egomotion is to be controlled based on vision, the body carrying the camera(s) may be subject to strong perturbations, which cannot be predicted, in general.

Pictorial vision (single image interpretation)	-------	***Motion vision***
Surveillance detection, inspection (prey)	-------	***Motion control*** (predator)
	[hybrid systems]	
Monocular *motion stereo*	-------	Bin- (multi-) ocular stereo
Passive	-------	***Active:*** *fixation type* *inertially stabilized,* *attention focused*
2-D shape	-------	***Spatial interpretation***
Off-line	-------	***Real-time***
Monochrome	-------	Color vision
Intensity	-------	Range

Figure 1.1. Types of vision systems and vision tasks

In cases with large rotational rates, motion blur may prevent image evaluation at all; also, due to the delay time introduced by handling and interpreting the large data rates in vision, stable control of the vehicle may no longer be possible.

Biological systems have developed close cooperation between inertial and optical sensor data evaluation for handling this case; this will be discussed to some detail and applied to technical vision systems in several chapters of the book. Also from biologists stems the differentiation of vision systems into “prey” and “predator” systems. The former strive to cover a large simultaneous field of view for detecting predators sufficiently early and approaching from any direction possible. Predators move to find prey, and during the final approach as well as in pursuit they have to estimate their position and speed relative to the dynamically moving prey quite accurate to succeed in a catch. Stereovision and high resolution in the direction of motion provides advantages, and nature succeeded in developing this combination in the vertebrate eye.

Once active gaze control is available, feedback of rotational rates measured by inertial sensors allows compensating for rotational disturbances on the own body just by moving the eyes (reducing motion blur), thereby improving their range of applicability. Fast moving targets may be tracked in smooth pursuit, also reducing motion blur for this special object of interest; the deterioration of recognition and tracking of other objects of less interest are accepted.

Since images are only in two dimensions, the 2-D framework looks most natural for image interpretation. This may be true for almost planar objects viewed approximately normal to their plane of appearance, like a landscape in a bird's-eye view. On the other hand, when a planar surface is viewed with the optical axis almost parallel to it from an elevation slightly above the ground, the situation is quite different. In this case, each line in the image corresponds to a different distance on the ground, and the same 3-D object on the surface looks quite different in size according to where it appears in the image. This is the reason why homogeneously distributed image processing by vector machines, for example, does have a hard time in showing its efficiency; locally adapted methods in image regions seem much more promising in this case and have proven their superiority. Interpreting image sequences in 3-D space with corresponding knowledge bases right from the beginning allows easy adaptation to range differences for single objects. Of course, the analysis of situations encompassing several objects at various distances now has to be done on a separate level, building on the results of all previous steps. This has been one of the driving factors in designing the architecture for the Third-generation "expectation-based, multi-focal saccadic" (EMS) vision system described in this book. This corresponds to recent findings in well-developed biological systems where for image processing and action planning based on the results of visual perception, different areas light up in magnetic resonance images [Talati, Hirsch 2005].

Understanding motion processes of 3-D objects in 3-D space while the body carrying the cameras also moves in 3-D space, seems to be one of the most difficult tasks in real-time vision. Without the help of inertial sensing for separating egomotion from relative motion, this can hardly be done successfully, at least in dynamic situations.

Direct range measurement by special sensors such as radar or laser range finders (LRF) would alleviate the vision task. Because of their relative simplicity and low demand of computing power, these systems have found relatively widespread application in the automotive field. However, with respect to resolution and flexibility of data exploitation as well as hardware cost and installation volume required, they have much less potential than passive cameras in the long run with computing power available in abundance. For this reason, these systems are not included in this book.

1.2 Why Perception and Action?

For technical systems which are intended to find their way on their own in an ever changing world, it is impossible to foresee every possible event and to program all required capabilities for appropriate reactions into its software from the beginning. To be flexible in dealing with situations actually encountered, the system should have perceptual and behavioral capabilities which it may expand on its own in response to new requirements. This means that the system should be capable of judging the value of control outputs in response to measured data; however, since outputs of control affect state variables over a certain amount of time, ensuing time histories have to be observed and a temporally deeper understanding has to be developed. This is exactly what is captured in the "dynamic models" of systems theory (and what biological systems may store in neuronal delay lines).

Also, through these time histories, the ground is prepared for more compact "frequency domain" (integral) representations. In the large volume of literature on linear systems theory, time constants T as the inverse of eigenvalues of first-order system components, as well as frequency, damping ratio, and relative phase as characteristic properties of second-order components are well known terms for describing temporal characteristics of processes, *e.g.*, [Kailath 1980]. In the physiological literature, the term "temporal Gestalt" may even be found [Ruhnau 1994a, b], indicating that temporal shape may be as important and characteristic as the well known spatial shape.

Usually, control is considered an output resulting from data analysis to achieve some goal. In a closed-loop system, where one of its goals is to adapt to new situations and to act autonomously, control outputs may be interpreted as questions asked with respect to real-world behavior. Dynamic reactions are now interpreted to better understand the behavior of a body in various states and under various environmental conditions. This opens up a new avenue for signal interpretation: beside its use for state control, it is now also interpreted for system identification and modeling, that is, learning about its temporal behavioral characteristics.

In an intelligent autonomous system, this capability of adaptation to new situations has to be available to reduce dependence on maintenance and adaptation by human intervention. While this is not yet state of the art in present systems, with the computing power becoming available in the future, it clearly is within range. The methods required have been developed in the fields of system identification and adaptive control.

The sense of vision should yield sufficient information about the near and farther environment to decide when state control is not so important and when more emphasis may be put on system identification by using special control inputs for this purpose. This approach also will play a role when it comes to defining the notion of a "self" for the autonomous vehicle.

1.3 Why Perception and Not Just Vision?

Vision does not allow making a well-founded decision on absolute inertial motion when another object is moving close to the ego-vehicle and no background can be seen in the field of view (known to be stationary). Inertial sensors like accelerometers and angular rate sensors, on the contrary, yield the corresponding signals for the body they are mounted on; they do this practically without any delay time and at high signal rates (up to the kHz range).

Vision needs time for the integration of light intensity in the sensor elements (33 1/3, respectively, 40 ms corresponding to the United States or European standard), for frame grabbing and communication of the (huge amount of) image data, as well as for feature extraction, hypothesis generation, and state estimation. Usually, three to five video cycles, that are 100 to 200 ms, will have passed until a control output derived from vision will hit the real world. For precise control of highly dynamic systems, this time delay has to be taken into account.

Since perturbations should be counteracted as soon as possible, and since visually measurable results of perturbations are the second integral of accelerations with corresponding delay times, it is advisable to have inertial sensors in the system for early pickup of perturbations. Because long-term stabilization may be achieved using vision, it is not necessary to resort to expensive inertial sensors; on the contrary, when jointly used with vision, inexpensive inertial sensors with good properties for the medium- to high-frequency part are sufficient as demonstrated by the vestibular systems in vertebrates.

Accelerometers are able to measure rather directly the effects of most control outputs; this alleviates system identification and finding the control outputs for reflex-like counteraction of perturbations. Cross-correlation of inertial signals with visually determined signals allows temporally deeper understanding of what in the natural sciences is called “time integrals” of input functions.

For all these reasons, the joint use of visual and inertial signals is considered mandatory for achieving efficient autonomously mobile platforms. Similarly, if special velocity components can be measured easily by conventional devices, it does not make sense to try to recover these from vision in a “purist” approach. These conventional signals may alleviate perception of the environment considerably since the corresponding sensors are mounted onto the body in a fixed way, while in vision the measured feature values have to be assigned to some object in the environment according to just visual evidence. *There is no constantly established link for each measurement value in vision as is the case for conventional sensors.*

1.4 What are Appropriate Interpretation Spaces?

Images are two-dimensional arrays of data; the usual array size today is from about 64 × 64 for special “vision” chips to about 770 × 580 for video cameras (special larger sizes are available but only at much higher cost, *e.g.*, for space or military applications). A digitized video data stream is a fast sequence of these images with data rates up to ~ 11 MB/s for black and white and up to three times this amount for color.

Frequently, only fields of 320 × 240 pixels (either only the odd or the even lines with corresponding reduction of the resolution within the lines) are being evaluated because of computing power missing. This results in a data stream per camera of about 2 MB/s. Even at this reduced data rate, the processing power of a single microprocessor available today is not yet sufficient for interpreting several video signals in parallel in real time. High-definition TV signals of the future may have up to 1080 lines and 1920 pixels in each line at frame rates of up to 75 Hz; this corresponds to data rates of more than 155 MB/s. Machine vision with this type of resolution is way out in the future.

Maybe, uniform processing of entire images is not desirable at all, since different objects will be seen in different parts of the images, requiring specific image processing algorithms for efficient evaluation, usually. Very often, lines of discontinuity are encountered in images, which should be treated with special methods differing essentially from those used in homogeneous parts. Object- and situation-dependent methods and parameters should be used, controlled from higher evaluation levels.

The question thus is, whether any basic feature extraction should be applied uniformly over the entire image region. In biological vision systems, this seems to be the case, for example, in the striate cortex (V1) of vertebrates where oriented edge elements are detected with the help of corresponding receptive fields. However, vertebrate vision has nonhomogeneous resolution over the entire field of view. Foveal vision with high resolution at the center of the retina is surrounded by receptive fields of increasing spread and a lower density of receptors per unit of area in the radial direction.

Vision of highly developed biological systems seems to ask three questions, each of which is treated by a specific subsystem:

1. Is there something of special interest in a wide field of view?
2. What is it precisely, that attracted interest in question one? Can the individual object be characterized and classified using background knowledge? What is its relative state “here and now”?
3. What is the situation around me and how does it affect optimal decisions in behavior for achieving my goals? For this purpose, a relevant collection of objects should be recognized and tracked, and the likely future behavior should be predicted.

To initialize the vision process at the beginning and to detect new objects later on, it is certainly an advantage to have a bottom-up detection component available all over the wide field of view. Maybe, just a few algorithms based on coarse resolution for detecting interesting groups of features will be sufficient to achieve this goal. The question is, how much computing effort should be devoted to this bottom-up component compared to more elaborate, model based top-down components for objects already detected and being tracked. Usually, single objects cover only a small area in an image of coarse resolution.

To answer question 2 above, biological vision systems direct the foveal area of high resolution by so-called saccades, which are very fast gaze direction changes with angular rates up to several hundred degrees per second, to the group of features arousing most interest. Humans are able to perform up to five saccades per second with intermediate phases of smooth pursuit (tracking) of these features, indicating a very dynamic mode of perception (time-sliced parallel processing). Tracking can be achieved much more efficiently with algorithms controlled by prediction according to some model. Satisfactory solutions may be possible only in special task domains for which experience is available from previous encounters.

Since prediction is a very powerful tool in a world with continuous processes, the question arises: What is the proper framework for formulating the continuity conditions? Is the image plane readily available as plane of reference? However, it is known that the depth dimension in perspective mapping has been lost completely: All points on a ray have been mapped into a single point in the image plane, irrespective of their distance, which has been lost. Would it be better to formulate all continuity conditions in 3-D physical space and time? The corresponding models are available from the natural sciences since Newton and Leibnitz have found that differential equations are the proper tools for representing these continuity conditions in generic form; over the last decades, simulation technology has provided the methods for dealing with these representations on digital computers.

In communication technology and in the field of pattern recognition, video processing in the image plane may be the best way to go since no understanding of the content of the scene is required. However, for orienting oneself in the real world through image sequence analysis, early transition to the physical interpretation space is considered highly advantageous because it is in this space that occlusions become easily understandable and motion continuity persists. Also, it is in this space that inertial signals have to be interpreted and that integrals of accelerations yield 3-D velocity components; integrals of these velocities yield the corresponding positions and angular orientations for the rotational degrees of freedom. Therefore, for visual dynamic scene understanding, images are considered intermediate carriers of data containing information about the spatiotemporal environment. To recover this information most efficiently, all internal modeling in the interpretation process is done in 3-D space and time, and the transition to this representation should take place as early as possible. Knowledge for achieving this goal is specific to single objects and the generic classes to which they belong. Therefore, to answer question 2 above, specialist processes geared to classes of objects and individuals of these classes observed in the image sequence should be designed for direct interpretation in 3-D space and time.

Only these spatiotemporal representations then allow answering question 3 by looking at these data of all relevant objects in the near environment for a more extended period of time. To be able to understand motion processes of objects more deeply in our everyday environment, a distinction has to be made between classes of objects. Those obeying simple laws of motion from physics are the ones most easily handled (*e.g.*, by some version of Newton’s law). Light objects, easily moved by stochastically appearing (even light) winds become difficult to grasp because of the variable properties of wind fields and gusts.

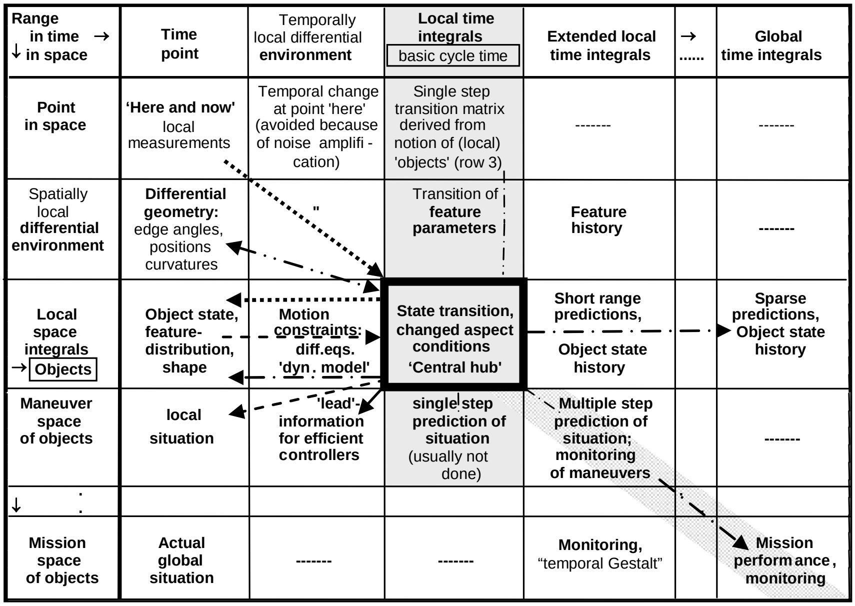
Another large class of objects – with many different subclasses – is formed by those able to sense properties of their environment and to initiate movements on their own, based on a combination of the data sensed and background knowledge internally stored. These special objects will be called **subjects**; all animals including humans belong to this (super-) class as well as autonomous agents created by technical means (like robots or autonomous vehicles). The corresponding subclasses are formed by combinations of perceptual and behavioral capabilities and, of course, their shapes. Beside their shapes, individuals of subclasses may be recognized also by stereotypical motion patterns (like a hopping kangaroo or a winding snake).

Road vehicles (independent of control by a human driver or a technical subsystem) exhibit typical behaviors depending on the situation encountered. For example, they follow lanes and do convoy driving, perform lane changes, pass other vehicles, turn off onto a crossroad or slow down for parking. All of the maneuvers mentioned are well known to human drivers, and they recognize the intention of performing one of those by its typical onset of motion over a short period of time. For example, a car leaving the center of its lane and moving consistently toward the neighboring lane is assumed to initiate a lane change. If this occurs within the safety margin in front, egomotion should be adjusted to this (improper) behavior of other traffic participants. This shows that recognition of the intention of other subjects is important for a defensive style of driving. This cannot be recognized without knowledge of temporally extended maneuvers and without observing behavioral patterns of subjects in the environment. Question 3 above, thus, is not answered by interpreting image patterns directly but by observing symbolic representations resulting as answers to question 2 for a number of individual objects/subjects over an extended period of time.

Simultaneous interpretation of image sequences on multiple scales in 3-D space and time is the way to satisfy all requirements for safe and goal-oriented behavior.

1.4.1 Differential Models for Perception "Here and Now"

Experience has shown that the simultaneous use of differential and integral models on different scales yields the most efficient way of data fusion and joint data interpretation. Figure 1.2 shows in a systematic fashion the interpretation scheme developed. Each of the axes is subdivided into four scale ranges. In the upper left corner the point "here and now" is shown as the point where all interaction with the real world takes place. The second scale range encompasses the local (as opposed to global) environment which allows introducing new differential concepts compared to the pointwise state. Local embedding, with characteristic properties



Range in time → ↓ in space	Time point	Temporally local differential environment	Local time integrals basic cycle time	Extended local time integrals	→	Global time integrals
Point in space	'Here and now' local measurements	Temporal change at point 'here' (avoided because of noise amplifi - cation)	Single step transition matrix derived from notion of (local) 'objects' (row 3)	-------		-------
Spatially local differential environment	Differential geometry: edge angles, positions curvatures	"	Transition of feature parameters	Feature history		-------
Local space integrals → Objects	Object state, feature-distribution, shape	Motion constraints: diff.eqs. 'dyn. model'	State transition, changed aspect conditions 'Central hub'	Short range predictions, Object state history		Sparse predictions, Object state history
Maneuver space of objects	local situation	'lead'-information for efficient controllers	single step prediction of situation (usually not done)	Multiple step prediction of situation; monitoring of maneuvers		-------
↓ :						
Mission space of objects	Actual global situation	-------	-------	Monitoring, "temporal Gestalt"		Mission performance, monitoring

Figure 1.2. Multiple interpretation scales in space and time for dynamic perception. Vertical axis: 3-D space; horizontal axis: time

such as spatial or temporal change rates, spatial gradients, or directions of extreme values such as intensity gradients are typical examples.

These differentials have shown to be powerful concepts for representing knowledge about physical properties of classes of objects. Differential equations represent the natural mathematical element for coding knowledge about motion processes in the real world. With the advent of the Kalman filter [Kalman 1960], they have become the key element for obtaining the best state estimate of the variables describing the system, based on recursive methods implementing a least-squares model fit. Real-time visual perception of moving objects is hardly possible without this very efficient approach.

1.4.2 Local Integrals as Central Elements for Perception

Note that the precise definition of what is local depends on the problem domain investigated and may vary in a wide range. The third column and row in Figure 1.2 are devoted to “local integrals”; this term again is rather fuzzy and will be defined more precisely in the task context. On the timescale, it means the transition from analog (continuous, differential) to digital (sampled, discrete) representations. In the spatial domain, typical local integrals are rigid bodies, which may move as a unit without changing their 3-D shape.

These elements are defined such that the intersection in field (3, 3) in Figure 1.2 becomes the central hub for data interpretation and data fusion: it contains the individual objects as units to which humans attach most of their knowledge about the real world. Abstraction of properties has lead to generic classes which allow subsuming a large variety of single cases into one generic concept, thereby leading to representational efficiency.

1.4.2.1 Where is the Information in an Image?

It is well known that information in an image is contained in local intensity changes: A uniformly gray image has only a few bits of information, namely, (1) the gray value and (2) uniform distribution of this value over the entire image. The image may be completely described by three bytes, even though the amount of data may be about 400 000 bytes in a TV frame or even 4 MB (2k × 2k pixels). If there are certain areas of uniform gray values, the boundary lines of these areas plus the internal gray values contain all the information in the image. This object in the image plane may be described with much less data than the pixel values it encompasses.

In a more general form, image areas defined by a set of properties (shape, texture, color, joint motion, *etc.*) may be considered image objects, which originated from 3-D objects by perspective mapping. Due to the numerous aspect conditions, which such an object may adopt relative to the camera, its potential appearances in the image plane are very diverse. Their representation will require orders of magnitude more data for an exhaustive description than its representation in 3-D space plus the laws of perspective mapping, which are the same for all objects. Therefore, an object is defined by its 3-D shape, which may be considered a local spatial integral of its differential geometry description in curvature terms. Depending on the task at hand, both the differential and the integral representation, or a combination of both may be used for visual recognition. As will be shown for the example of road vehicle guidance, the parallel use of these models in different parts of the overall recognition process and control system may be most efficient.

1.4.2.2 To Which Units Do Humans Affix Knowledge?

Objects and object classes play an important role in human language and in learning to understand “the world”. This is true for their appearance at one time, and also for their motion behavior over time.

On the temporal axis, the combined use of differential and integral models may allow us to refrain from computing optical flow or displacement vector fields, which are very compute-intensive and susceptible to noise. Because of the huge amount of data in a single image, this is not considered the best way to go, since an early transition to the notion of physical objects or subjects with continuity conditions in 3-D space and time has several advantages: (1) it helps cut the amount of data required for adequate description, and (2) it yields the proper framework for applying knowledge derived from previous encounters (dynamic models, stereotypical control maneuvers, *etc.*). For this reason, the second column in Figure 1.2 is avoided intentionally in the 4-D approach. This step is replaced by the well-known observer techniques in systems dynamics (Kalman filter and derivatives, Luenberger observers). These recursive methods reconstruct the time derivatives of state variables by prediction error feedback and knowledge about the dynamic behavior of the object and (for the Kalman filter) of the statistical properties of the system (dubbed “plant” in systems dynamics) and of the measurement processes. The stereotypical behavioral capabilities of subjects in different situations form an important part of the knowledge base.

Two distinctly different types of “local temporal integrals” are used widely: Single step integrals for video sampling and multiple step (local) integrals for maneuver understanding. Through the imaging process, the analog motion process in the real world is made discrete along the time axis. By forming the (approximate, since linearized) integrals, the time span of the analog video cycle time (33 1/3 ms in the United States and 40 ms in Europe, respectively, half these values for the fields) is bridged by discrete transition matrices from kT to $(k + 1)T$, k = running index.

Even though the intensity values of each pixel are integrals over the full range or part of this period, they are interpreted as the actually sampled intensity value at the time of camera readout. Since all basic interpretations of the situation rest on these data, control output is computed newly only after this period; thus, it is constant over the basic cycle time. This allows the analytical computation of the corresponding state transitions, which are evaluated numerically for each cycle in the recursive estimation process (Chapter 6); these are used for state prediction and intelligent control of image feature extraction.

1.4.3 Global Integrals for Situation Assessment

More complex situations encompassing many objects or missions consisting of sequences of mission elements are represented in the lower right corner of Figure 1.2. Again, how to best choose the subdivisions and the absolute scales on the time axis or in space depends very much on the problem area under study. This will be completely different for a task in manufacturing of micro-systems compared to one in space flight. The basic principle of subdividing the overall task, however, may be according to the same scheme given in Figure 1.2, even though the technical elements used may be completely different.

On a much larger timescale, the effect of entire feed-forward control time histories may be predicted which have the goal of achieving some special state changes or transitions. For example, lane change of a road vehicle on a freeway, which may take 2 to 10 seconds in total, may be described as a well-structured sequence of control outputs resulting in a certain trajectory of the vehicle. At the end of the maneuver, the vehicle should be in the neighboring lane with the same state variables otherwise (velocity, lateral position in the lane, heading). The symbol “lane change”, thus, stands for a relatively complex maneuver element which may be triggered from the higher levels on demand by just using this symbol (maybe together with some parameters specifying the maneuver time and, thereby, the maximal lateral acceleration to be encountered). Details are discussed in Section 3.4.

These “maneuver elements”, defined properly, allow us to decompose complex maneuvers into stereotypical elements which may be pieced together according to the actual needs; large sections of these missions may be performed by exploiting feedback control, such as lane following and distance keeping for road vehicles. Thereby, scales of distances for entire missions depend on the process to be controlled; these will be completely different for “autonomously guided vehicles” (AGVs) on the factory floor (hundreds of meters) compared to road vehicles (tens of km) or even aircraft (hundreds or thousands of km).

The design of the vision system should be selected depending on the task at hand (see next section).

1.5 What Type of Vision System Is Most Adequate?

For motion control, due to inertia of a body, the actual velocity vector determines where to look to avoid collisions with other objects. Since lateral control may be applied to some extent and since other objects and subjects may have a velocity vector of their own, the viewing range should be sufficiently large for detecting all possible collision courses with other objects. Therefore, the simultaneous field of view is most critical nearby.

On the other hand, if driving at high speed is required, the look-ahead range should be sufficiently large for reliably detecting objects at distances which allow safe braking. At a speed of 30 m/s (108 km/h or about 65 mph), the distance for braking [with a deceleration level of 0.4 Earth gravity g (9.81 m/s^2, that is $a_x \approx -4$

m/s²) and with 0.5 seconds reaction time] is 15 + 113 = 128 m. For half the magnitude in deceleration (− 2 m/s^2, *e.g.,* under unfavorable road conditions) the braking distance would be 240 m.

Reliable distance estimation for road vehicles occurs under mapping conditions with at least about 20 pixels on the width of the vehicle (typically of about 2 m in dimension). The total field of view of a single camera at a distance of 130 m, where this condition is satisfied, will be about 76 m (for ~ 760 pixel per line). This corresponds to an aperture angle of ~ 34°. This is certainly not enough to cover an adequate field of view in the near range. Therefore, at least a bifocal camera arrangement is required with two different focal lengths (see Figure 1.3).

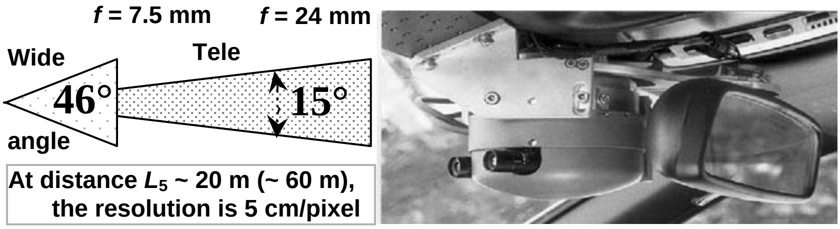


Figure 1.3. Bifocal arrangement of miniature TV–cameras on a pan platform in front of the rear view mirror of test vehicles **VaMP** and **VITA 2**, Prometheus, 1994. Left: Fields of view and ranges (schematically), right: System realized in VaMP

For a rather flexible high performance “technical eye” a trifocal camera arrangement as shown in Figure 1.4 is recommended. The two wide-angle CCD-cameras with focal length of 4 to 6 mm and with divergent optical axes do have a central range of overlapping image areas, which allows stereo–interpretation nearby. In total, a field of view of about 100 to 130 degrees can be covered; this allows surveying about one–third of the entire panorama.

The mild telecamera with three to four times the focal length of the wide-angle one should be a three–chip color camera for more precise object recognition. Its field of view is contained in the stereo field of view of the wide-angle cameras such that trinocular stereointerpretation becomes possible [Rieder 1996].

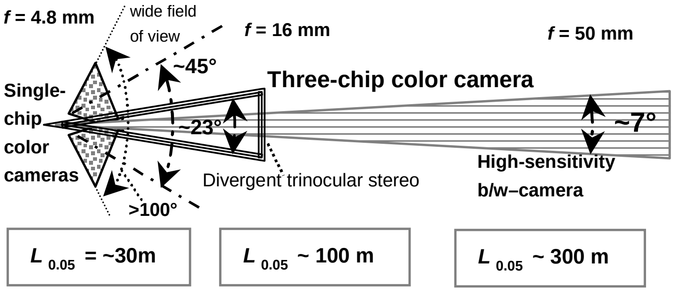


Figure 1.4. Trifocal camera arrangement with wide field of view

To detect objects in special areas of interest far away, a camera with a third focal length (again with a factor of 3 to 4 relative to the mild telelens), and the field of view within that of the mild telecamera should be added (see Figure 1.4). This camera may be chosen to be especially light-sensitive; black-and-white images may be sufficient to limit the data rate. The focal length ratio of 4 does have the advantage that the coarser image represents the same scene at a resolution corresponding to the second pyramidal stage of the finer one.

This type of sensor combination is ideally suited for active viewing direction control: the coarse resolution, large simultaneous field of view allows discovering objects of possible interest in a wide area, and a viewing direction change will bring this object into the center of the images with higher resolution. Compared to a camera arrangement with maximal resolution in the same entire field of view, the solution shown has only 2 to 4 % the data rate. It achieves this in exchange for the need of fast viewing direction control and at the expense of delay times required to perform these gaze changes. Figure 1.5 gives an impression of the fields of view of this trifocal camera arrangement.

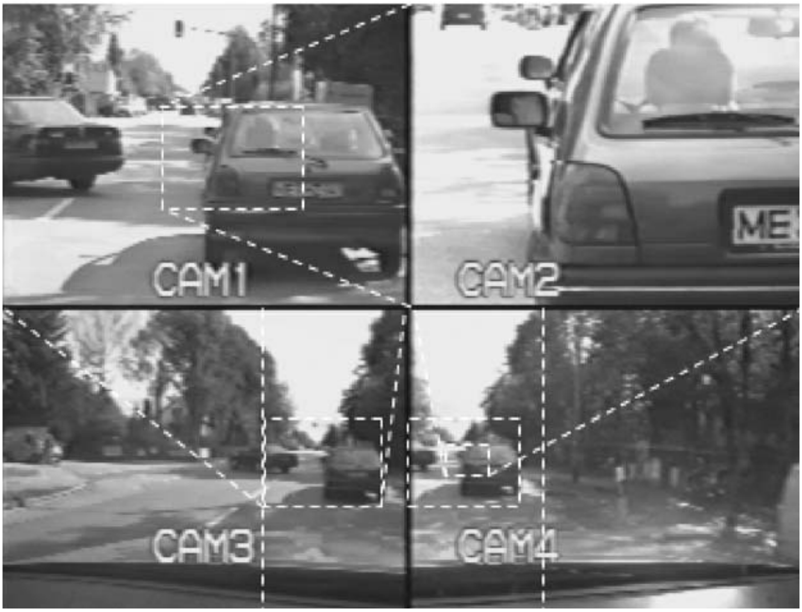


Figure 1.5. Fields of view of trifocal camera arrangement. Bottom: Two divergent wide angle cameras; top left: mild tele camera, top right: strong tele-camera. Dashed white lines show enlarged sections

The lower two wide-angle images have a central region of overlap marked by vertical white lines. To the left, the full road junction is imaged with one car coming out of the crossroad and another one just turning into the crossroad; the rear of this vehicle and the vehicle directly in front can be seen in the upper left image of the mild telecamera. This even allows trinocular stereo interpretation. The region marked in white in this mild teleimage is shown in the upper right as a full image of the strong telecamera. Here, letters on the license plate can be read, and it can be seen from the clearly visible second rearview mirror on the left-hand side that there is a second car immediately in front of the car ahead. The number of pixels per area on the same object in this image is one hundred times that of the wide-angle images.

For inertial stabilization of the viewing direction when riding over a nonsmooth surface or for aircraft flying in a turbulent air, an active camera suspension is needed anyway. The simultaneous use of almost delay-free inertial measurements (time derivatives such as angular rates and linear accelerations) and of images, whose interpretation introduces several tenths of a second delay time, requires extended representations along the time axis. There is no single time for which it is possible to make consistent sense of all data available. Only the notion of an "extended presence" allows arriving at an efficient invariant interpretation (in 4-D!). For this reason, the multifocal, saccadic vision system is considered to be the preferable solution for autonomous vehicles in general.

1.6 Influence of the Material Substrate on System Design: Technical vs. Biological Systems

Biological vision systems have evolved over millions of generations with the selection of the fittest for the ecological environment encountered. The basic neural substrate developed (carbon-based) may be characterized by a few numbers. The electrochemical units do have switching times in the millisecond (ms) range; the traveling speed of signals is in the 10 to 100 m/s range. Cross-connections between units exist in abundance (1000 to 10 000 per neuron). A single brain consists of up to 10^{11} of these units. The main processing step is summation of the weighted input signals which contain up to now unknown (multiple?) feedback loops [Handbook of Physiology 1984, 1987].

These systems need long learning times and adapt to new situations only slowly. In contrast, technical substrates for sensors and microprocessors (silicon-based) have switching times in the nanosecond range (a factor of 10^6 compared to biological systems). They are easily programmable and have various computational modes between which they can switch almost instantaneously; however, the direct cross-connections to other units are limited in number (one to six, usually) but may have very high bandwidth (in the hundreds of MB/s range).

While a biological eye is a very complex unit containing several types and sizes of sensors and computing elements, technical imaging sensors are rather simple up to now and mostly homogeneous over the entire array area. However, from television and computer graphics, it is well known that humans can interpret the images thus generated without problems in a natural way if certain standards are maintained.

In developing dynamic machine vision, two groups of thinking have formed: One tries to mimic biological vision systems on the silicon substrate available, and the other continues to build on the engineering platform developed in systems– and computer science.

A few years ago, many systems were investigated with single processors devoted to single pixels (Connection Machine [Hillis 1985, 1992], Content-Addressable Associative Parallel Processors (CAAPP) [Scudder, Weems 1990] and others). The trend now clearly is toward more coarsely granulated parallel architectures. Since a single microprocessor on the market at the turn of the century is capable of performing about 10^9 instructions per second, this means in excess of 2000 instructions per pixel of a 770 × 525 pixel image. Of course, this should not be confused with information processing operations. For the year 2010, general-purpose PC processors are expected to have a performance level of about 10^{11} instructions per second.

On the other hand, the communication bandwidths of single channels will be so high, that several image matrices may be transferred at a sufficiently high rate to allow smooth recognition and control of motion processes. (One should refrain from video norms, presently dominating the discussion, once imaging sensors with digital output are in wide use.) Therefore, there is no need for more elaborate data processing on the imaging chip except for ensuring sufficiently high intensity dynamics. Technical systems do not have the bandwidth problems, which may have forced biological systems to do extensive data preprocessing near the retina (from 120 million light sensitive elements in the retina to 1.2 million nerves leading to the lateral geniculate nucleus in humans).

Interesting studies have been made at several research institutions which tried to exploit analog data processing on silicon chips [Koch 1995]; future comparisons of results will have to show whether the space needed on the chip for this purpose can be justified by the advantages claimed.

The mainstream development today is driven by commercial TV for the sensors and by personal computers and games for the processors. With an expected increase in computing power of one order of magnitude every 4 to 5 years over the next decade, real-time machine vision will be ready for a wide range of applications using conventional engineering methods as represented by the 4-D approach.

A few (maybe a dozen) of these processors will be sufficient for solving even rather complex tasks like ground and air vehicle guidance; dual processors on a single chip are just entering the market. It is the goal of this monograph to make the basic methods needed available to a wide public for efficient information extraction from huge data streams.

1.7 What Is Intelligence? A Practical (Ecological) Definition

The sensors of complex autonomous biological or technical systems yield an enormous data rate containing information about both the state of the vehicle body relative to the environment and about other objects or subjects in the environment. It is the task of an intelligent information extraction (data interpretation) system to quickly get rid of as many data as possible, however simultaneously, to retain all of the essential information for the task to be solved. Essential information is geared to task domains; however, complex systems like animals and autonomous vehicles