Robotics and Autonomous Systems 8 (1991) 65–91
North-Holland

Understanding object motion: Recognition, learning and spatiotemporal reasoning

Michael Mohnhaupt and Bernd Neumann

Universität Hamburg, Fachbereich Informatik, Bodenstedtstr. 16, D-2000 Hamburg 50, Germany

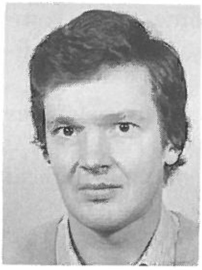
Abstract

Mohnhaupt, M. and Neumann, B., Understanding object motion: Recognition, learning and spatiotemporal reasoning. Robotics and Autonomous Systems, 8 (1991) 65–91.

Modeling object motion is an important taks for intelligent systems. There are several cognitive tasks related to object motion, e.g., concept formation, recognition and prediction. In this paper, we present an integrated hybrid representational system for object motion, in which different modes of representation are exploited. The representational system includes a propositional qualitative long-term representation, and an analog quantitative short-term representation, which is instantiated on demand. We show how several tasks related to object motion can be solved: (1) Learning of object motion, from single observations towards propositional models. The central idea is to exploit a rich set of perceptual primitives for several learning tasks. (2) Recognition of spatiotemporal events based on qualitative predicates, and subsequent natural language description, and (3) several important aspects of spatiotemporal reasoning.

Keywords: Pictorial representations; Propositional representations; Spatiotemporal events; Learning in pictorial representations; Event recognition; Spatiotemporal reasoning; High-level vision.

Michael Mohnhaupt finished his diploma degree in computer science in 1985 at the University of Hamburg. From 1985 to 1986 he studied at the University of Toronto based on a scholarship from DAAD. Since September 1986 he has been a research assistant at the University of Hamburg. He is currently working on his Ph.D. in the area of high-level vision and knowledge representation.



Bernd Neumann studied Electrical Engineering at the University of Darmstadt (Diploma 1967) and Information Theory at MIT/USA (M.S. 1968, Ph.D. 1971). He began his university activities as a lecturer for Computer Science at the University of Hamburg. Since 1982 he is Professor at the Department of Computer Science, since 1986 he has been head of the research group 'Cognitive Systems' at the University of Hamburg. He is also head of the Artificial Intelligence Laboratory in Hamburg which offers application-oriented consultation and cooperation. His research area is Artificial Intelligence with special interests in Computer Vision, Robotics and Expert Systems.

1. Introduction

The understanding of object motion is crucial for many of the sophisticated tasks commonly required of biological and robotic systems. Naturally, vision is one of the principal tools that provides an observer with the necessary information; and not surprisingly, a substantial portion of the brain is dedicated to visual perception. In addition, it seems clear that our perceptual capabilities are complemented by cognition and spatiotemporal reasoning. For example, we can expect the perceptual system to extract motion information and to generate an internal description of path and structure of an object. It also seems likely that prototypical classes of motion seen repeatedly within the environment are abstracted to facilitate path planning and recognition of moving objects, and to predict object trajectories well beyond the simple extrapolation

0921-8830/91/$03.50

that might be expected of the perceptual system. As a result the perceptual system would be faster and more robust. It would have a deeper understanding and therefore predictive power of spatiotemporal behavior.

Although the general goal of obtaining motion representation with predictive power seems indisputable, there remain difficult questions concerning the kind of representations suitable for the diverse tasks related to object motion and the nature of the processes which go along with the representations. There are severe constraints on possible representations due to temporal limitations for processing and due to the complexity of many perceptual and cognitive tasks for biological and artificial systems [1] (see e.g. [19,42]). In our view this necessitates the use of different specialized representations including different modes of processing.

In this paper we present a fairly general framework for understanding object motion. Our approach is influenced by work on knowledge representation in the area of Artificial Intelligence and by work on possible cognitive representations in humans, but we do not claim cognitive adequacy with respect to humans. Nevertheless, the human cognitive system is a useful model for developing artificial cognitive systems, in particular, if no solid theories are available for the domain of interest. Consequently, many aspects of our approach do not contradict theories on cognitive representations in humans and might influence possible models.

We employ both, a propositional qualitative and an analogical [2] quantitative representation to solve several important tasks related to object motion. The propositional representation supports recognition using logic-based reasoning and is suited for natural language communication. It is used as long-term representation. The analogical representation (called spatiotemporal buffer) is essential for perceiving and visualizing object motion as well as for learning and important aspects of spatiotemporal reasoning. [3]

It is evident that the usefulness of the analogical representation and the local processes working on it mainly derives from the fact that this representation is specialized for concrete visual data, given through perceptual processes or instantiated from models in long-term memory. For example, important spatiotemporal relations are explicit in this representation and therefore easily accessible, and important physical constraints are intrinsically coded. Also, substantial use can be made of (subsymbolic) local parallel processing as will be shown. Moreover, cognitive processes merge smoothly with perceptual processes within the analogical representation. It provides a shared representation for bottom-up processes (e.g. building models from concrete visual data), for top-down processes (e.g. top-down guided low-level vision analysis), as well as for processes which rely on information from both directions (e.g. adjusting generic models to perceptual data).

In addition, the hybrid framework supports learning of object motion, which is a fundamental concern for any intelligent system. For example, there is a natural transition from concrete observations of object motion recorded in the spatiotemporal buffer to accumulated experience resulting in generic event models and propositional descriptions.

Although a solid representational theory is still lacking, the hybrid approach presented here suggests that different representations including different modes of reasoning are advantageous for important tasks in the area of object motion, and that an analogical representation including local processes plays a special part. There are similar results for other relevant spatiotemporal domains, including obstacle avoidance and reasoning about non-solid objects (see e.g. [13,41]), and for purely spatial problems, including scanning tasks, computing spatial relations, and reasoning about diagrams (see e.g. [17,18,20].

We choose object motions in traffic scenes as our experimental application domain. Here typical objects are cars and pedestrians etc., and typical

[1] The terms *perceptual* and *cognitive* refer in this paper to processes in biological and artificial information processing systems.

[2] In this paper a dimension of a representation is called *analogical* if the mapping from the modeled world into the modeling world preserves the inherent structure of this dimension (see [32,34]).

[3] Here, we only deal with those aspects of spatiotemporal reasoning which are concerned with the concrete visual world and abstractions thereof.

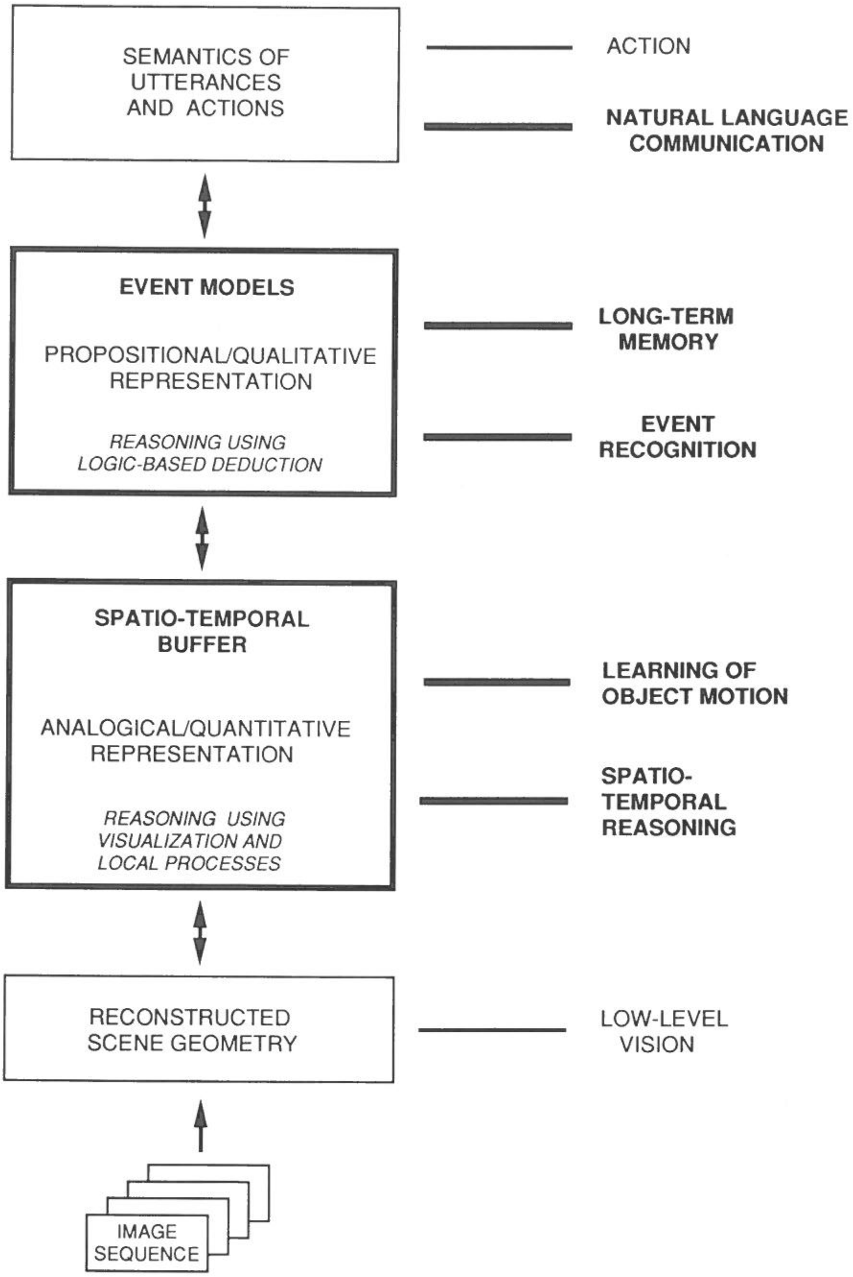

SEMANTICS OF UTTERANCES AND ACTIONS
EVENT MODELS
PROPOSITIONAL/QUALITATIVE REPRESENTATION
REASONING USING LOGIC-BASED DEDUCTION
SPATIO-TEMPORAL BUFFER
ANALOGICAL/QUANTITATIVE REPRESENTATION
REASONING USING VISUALIZATION AND LOCAL PROCESSES
RECONSTRUCTED SCENE GEOMETRY
IMAGE SEQUENCE
ACTION
NATURAL LANGUAGE COMMUNICATION
LONG-TERM MEMORY
EVENT RECOGNITION
LEARNING OF OBJECT MOTION
SPATIO-TEMPORAL REASONING
LOW-LEVEL VISION

Fig. 1. Understanding object motion. The tasks relevant for object motion (right column) and the main representations used (left column), including a brief characterization.

object motion includes 'turn-off' events, 'overtake' events, 'walk' events and similar meaningful spatiotemporal occurrences.

In Section 2 we outline: (1) the tasks related to object motion which will be attacked, (2) the hybrid representation of object motion, and (3) its associated processes. In addition, some motivation will be provided for subsequent sections.

2. A hybrid representation for object motion

Our understanding of motion has developed from research into distinct tasks which must be supported by adequate object motion representations. *Fig. 1* provides a framework for identifying these tasks.

Low-level vision. Beginning at the bottom, we have the task of low-level image analysis. This is, of course, the main task of computer vision. It includes the initial detection of motion up to the interpretation of motion phenomena in terms of changing object positions. Many subproblems of low-level vision are not solved but we assume for now that it provides a reconstruction of the time-varying 3D scene geometry as its output. Hence structure and spatiotemporal positions of objects are assumed to be known, but objects do not have to be identified at this level.

Long-term memory. As the next task we consider recording spatiotemporal information of object motion for long-term memory. This is clearly a desideratum for building representations with predictive power. For many reasons, including biological plausibility (e.g. [22]), we assume that a representation more compact than a quantitative analogical scene description must be computed for long-term memory of motion concepts. It should be qualitative in nature abstracting from quantitative spatiotemporal information, and can be thought of in terms of predicates or propositions describing the scene.

Recognition and natural language communication. Propositional representations are also useful for describing generic concepts, for example, models for characteristic object motions. We have investigated the use of such *event models* for recognition in earlier work (see [30,31]). An event model would describe the characteristic properties of object trajectories in traffic scenes such as 'overtaking', 'turning-off', 'crossing', etc. Similar models are used by Andre et al. [2] for event recognition of ongoing object motion in soccer games. In Tsotsos et al. [43] motion concepts are represented in frames connected through semantic networks. Propositional event models also proved useful for natural language communication about scenes. From instantiated event models we could fill case frames for natural language utterances and generate a coherent description. This work is summarized in Section 3.

Learning object motion. Another relevant task within the model outlined in *Fig. 1* is learning. Given an initially empty store of event models and a world full of motion phenomena, how can event models be generated to begin with?

In the ideal case, an intelligent system would come into existence equipped with a collection of a priori models for spatiotemporal events. This collection would be sufficiently rich to allow for efficient recognition and prediction over a wide range of environmental conditions and objects of interest. Our understanding of elementary physics would be one way to obtain such models. Based on the theoretical understanding of the environment we can build 'analytical' models, for example, to predict the path of a thrown ball. The analytical model might exploit knowledge of initial forces, the gravitation, the initial angle of elevation, the weight of the ball and so on. Such models are common in path planning and naive physics (e.g. [6,11]). Unfortunately, they do not, at present, generalize to the wide range of complex situations occurring in our environment. Moreover, the necessary initial conditions may never be available with sufficient accuracy, nor might these models lead to efficient reasoning.

As an alternative, a system could acquire models by gradually building up knowledge about the environment, observing objects and their interactions. The ability to learn allows for generality, adaptability and extensibility (e.g. [23]). It is also closely related to long-term memory because replicas of past experiences can be viewed as basic generic models with predictive power.

For a flexible system, models must be adaptable to new surroundings. This can be done, in principle, by abstracting experiences gained in one environment and carrying it over to another. For example, an abstracted representation of the motion of thrown baseballs can be used to predict the path of baseballs or an unknown object in similar situations.

In this paper we concentrate on the use of spatiotemporal event models that are learned from concrete observations and therefore reflect some typical behavior of certain objects within a given environment. Prototypical behavior can be represented by abstract trajectories based on the average over many single instances, or based on a single observed instance which is deemed to be sufficiently representative of other instances or to be of particular importance.

The accumulation of specific observations and the extraction of prototypical behavior takes place based on visualizations within an analogical representation, the *spatiotemporal buffer* (see *Fig. 1*). Prototypical trajectories, along with other relevant information discovered during the learning process (e.g. distribution of paths about the prototypical trajectory), are then encoded in a propositional and compact manner using *event models*. Details of the learning task will be described in Section 4.

Spatiotemporal reasoning. Another interesting task investigated within our framework is spatiotemporal reasoning, using the *spatiotemporal buffer*. Several aspects will be discussed in detail in Section 5.

Spatiotemporal reasoning, involving path planning, prediction, or the extraction of spatiotemporal relations, has been considered within propositional frameworks, using logic-based forms of reasoning and inference (see e.g. [11]). However, there are significant drawbacks in this sort of approach in terms of tractability (see [19] for examples). In particular, complexity may grow exponentially with the number of propositions. It may also be problematic to determine the facts relevant for specific spatiotemporal changes (frame problem). A similar problem concerns consistency and completeness of propositional descriptions. For example, a natural language description of a given scene should be consistent and complete with respect to some level of abstraction. Both is very difficult to maintain on a purely propositional level.

The approach considered here attempts to deal with prediction and path planning by visualizations and local processes using relevant scene information and experience-based models of spatiotemporal events in a common analogical representation. The analogical representation is essentially the same used for learning and accumulating experience, the short-term spatiotemporal buffer [4] instantiated by demand and initialized from a priori knowledge about the environment, the learned knowledge about typical trajectories, and current information provided by the perceptual processes. Thus, information is included from both cognitive and perceptual sources It is interesting to note that there is evidence for shared representational structures between cognitive and perceptual processes in humans (see e.g [9]).

Spatiotemporal reasoning using visualizations and local processes in an analogical buffer can be viewed as subsymbolic processing (see [41]). Like Steels we feel that certain reasoning tasks, especially those associated with space and time, are well suited for a subsymbolic mode of processing which may be controlled by a more abstract, symbolic level. On a subsymbolic level the processes can be constrained to yield physically plausible solutions. In particular, the relevant information (from experience or perceptual analysis) is explicitly loaded into the analogical buffer, after which the reasoning process inherently determines spatiotemporal interaction between objects. Although this approach is less general than one based on logic, it is hoped that it will yield tractable solutions with sufficient richness in the spatiotemporal domain. There are other results which support this idea. For example, Funt's work [12] on prediction of the collision of falling objects based on an array-like representation and recent interesting work by Gardin and Meltzer [13] on modeling the behavior of non-solid objects based on an analogical representation.

Another relevant reasoning problem arises when motion analysis and event recognition are carried out in the context of top-down constraints. [5] For example, if the only event of interest is whether or not a car approaches, how can this knowledge be put to use for efficient motion analysis? Visualizations are appropriate to express top-down information because they are closely related to visual

[4] Which is similar to the purely spatial buffer proposed by Kosslyn [17].

[5] Top-down information for visual processes is almost permanently available through intentions, expectations, a priori knowledge and the like.

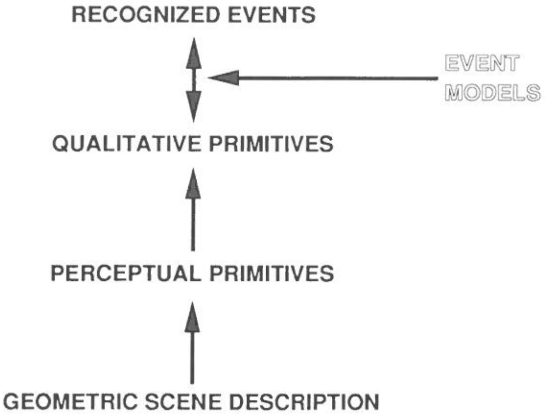


Fig. 2. Main processing stages for event recognition.

perceptions. Therefore they can easily be used to constrain visual processes.

Choosing an adequate representational scheme and mode of computation is a key problem for intelligent systems. There are few sound principles to favor one representation over another, and there is no coherent representational theory, although there are some interesting attempts (see e.g. [32]). It is instructive to view our approach from this perspective and to derive some general constraints on representations discussing the merits of different representations used in our work. A more detailed discussion of these issues can be found in Section 6.

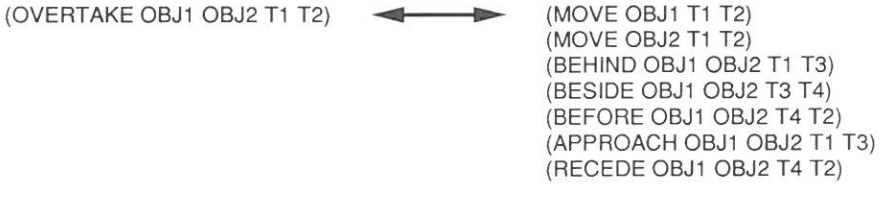
3. Recognition and verbalization

In this section we present representations and processes developed for the recognition and verbalization of events in street traffic scenes. An event is defined as a subset of the scene which can be described by a certain verb of locomotion, e.g. 'overtake'. A priori knowledge about event classes is provided by propositional event models. Event recognition based on such models is implemented by hierarchical matching which takes propositional primitives describing the current scene as input and generates instantiated events as output. The propositional primitives are qualitative predicates computed from a certain set of quantitative perceptual primitives which in turn are generated from the spatiotemporal scene description rendered by vision.

Fig. 2 shows the main processing stages for event recognition. A detailed description of event recognition can be found in Neumann [30]. This section reviews the type of motion representation developed for this purpose.

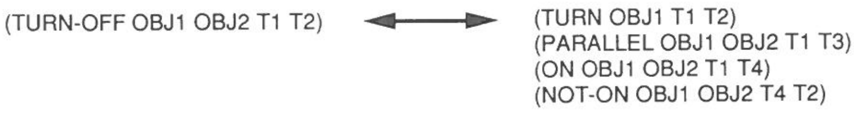
Event models have been tailored around verbs of locomotion to support verbalization of events. This naturally leads up to propositional models. *Fig. 3* shows the event model for 'overtake'.

It consists of a head which describes the event in terms of a proposition, and a body which contains premisses for the event to be true. The parameters following the predicate identifier are



```
(OVERTAKE OBJ1 OBJ2 T1 T2)   <——>   (MOVE OBJ1 T1 T2)
                                    (MOVE OBJ2 T1 T2)
                                    (BEHIND OBJ1 OBJ2 T1 T3)
                                    (BESIDE OBJ1 OBJ2 T3 T4)
                                    (BEFORE OBJ1 OBJ2 T4 T2)
                                    (APPROACH OBJ1 OBJ2 T1 T3)
                                    (RECEDE OBJ1 OBJ2 T4 T2)
```

Fig. 3. Event model for 'overtake'.



```
(TURN-OFF OBJ1 OBJ2 T1 T2)   <——>   (TURN OBJ1 T1 T2)
                                    (PARALLEL OBJ1 OBJ2 T1 T3)
                                    (ON OBJ1 OBJ2 T1 T4)
                                    (NOT-ON OBJ1 OBJ2 T4 T2)
```

Fig. 4. Event model for 'turn-off'.

variables which must be instantiated during event recognition. $T1$, $T2, \ldots$ denote time variables which mark the beginning and ending of time intervals.

Another example is 'turn-off' described in *Fig. 4*. Here OBJ1 is a vehicle and OBJ2 is a street.

Some of the predicates can be decomposed further, e.g. 'turn' can be 'left-turn' or 'right-turn'. Event models constitute a specialization hierarchy with the most general event ('exist') as a root and increasingly complex events built up by composition and specialization. In our project some 50 event models have been defined in this manner.

It is interesting to take a look at the primitives of this representational system. They have the following properties:

- Primitive predicates are durative, i.e. they describe time intervals where certain scene properties are continuously valid.
- Primitive predicates are qualitative by nature. They give rise to qualitative propositions about quantitative perceptual primitives derived from the scene.

The perceptual primitives required for this task are essentially object positions relative to reference positions and object orientations relative to reference orientations, each quantity as a function of time and including temporal derivations. There may be many interesting reference locations and orientations, hence the set of perceptual primitives may be quite large. Nevertheless, they constitute a well-defined perceptual basis for higher-level descriptions.

Basically, the purpose of primitive predicates is to derive interesting constancies among the perceptual primitives. For event recognition the following predications turned out to be necessary and sufficient:

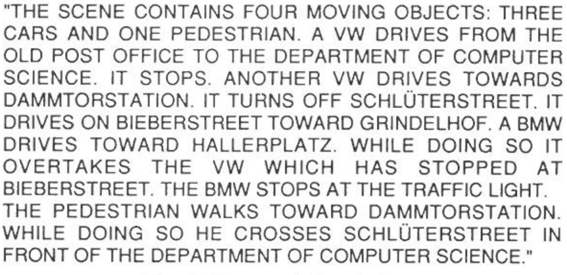
- Constant value
  (e.g. standing-still, moving straight, keeping constant velocity,...),
- Monotonicity
  (e.g. acceleration, turning, approach,...),
- Limited value set
  (e.g. parallel, close-to, beside, on,...)
- Greater/smaller
  (e.g. faster than usual,...).

Applying these predicates to suitably chosen perceptual primitives results in a limited set of primitive events which are the basis for the event hierarchy. We employed 19 primitive events for traffic scene descriptions.

Details about the event recognition procedure can be found in Neumann [30]. The procedure is similar to hierarchical matching as employed elsewhere in AI, e.g. for object recognition. It may be noteworthy that a constraint network has to be maintained to deal with temporal constraints arising from the time variables.

Once events have been found, there remain several nontrivial processing steps until a coherent natural language description can be generated. Typically, a large number of events are candidates for verbalization, hence there is the problem of choosing the most 'informative' event. In our work, the most special events are selected according to the specialization hierarchy. Second, there is the problem of ordering event-based utterances into a coherent description. Here a chronological order is followed for each single moving object. Third, temporal and locative phrases have to be generated, enough to give a complete scene description but also avoiding redundancy.

We have introduced the notion of 'anticipated visualization' as a guiding principle for generating an informative scene description. The idea is to simulate the hearer's understanding process as a means of guiding the generation process. For example, the decision to include a locative phrase in an utterance – say 'on the right side of Schlueterstreet' – should be made depending on whether or not the hearer needs this information to be able to visualize the scene. In our project, anticipated visualization has only been implemented implicitly in terms of a 'standard plan' for generating visualizable descriptions. *Fig. 5* shows a description (translated from German) generated for the synthetic scene shown in *Fig. 6*.



```
"THE SCENE CONTAINS FOUR MOVING OBJECTS: THREE
CARS AND ONE PEDESTRIAN. A VW DRIVES FROM THE
OLD POST OFFICE TO THE DEPARTMENT OF COMPUTER
SCIENCE. IT STOPS. ANOTHER VW DRIVES TOWARDS
DAMMTORSTATION. IT TURNS OFF SCHLÜTERSTREET. IT
DRIVES ON BIEBERSTREET TOWARD GRINDELHOF. A BMW
DRIVES TOWARD HALLERPLATZ. WHILE DOING SO IT
OVERTAKES THE VW WHICH HAS STOPPED AT
BIEBERSTREET. THE BMW STOPS AT THE TRAFFIC LIGHT.
THE PEDESTRIAN WALKS TOWARD DAMMTORSTATION.
WHILE DOING SO HE CROSSES SCHLÜTERSTREET IN
FRONT OF THE DEPARTMENT OF COMPUTER SCIENCE."
```

Fig. 5. Generated description.

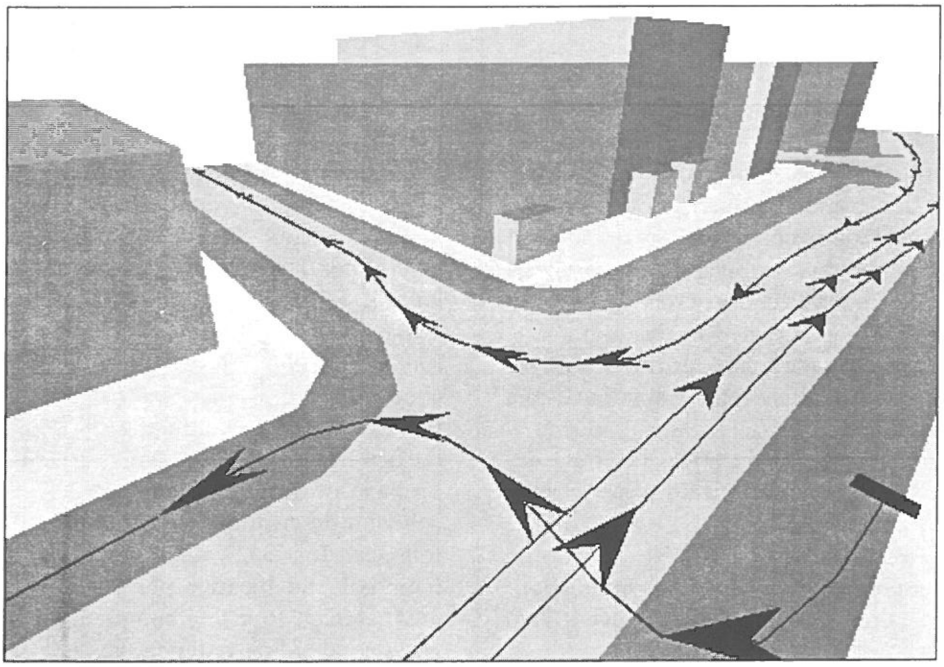


Fig. 6. Synthetic scene with 4 moving objects.

Thinking about anticipated visualization, we found out soon that propositional event descriptions alone, as might be recovered by the hearer from the natural language description, do not provide sufficient information to regenerate the scene in sufficient detail. To begin with, a natural language description is qualitative by nature, hence there are many possible quantitative scene descriptions fitting the same description. In face of this inherent ambiguity humans tend to settle for a 'typical' visualization. But how can typical visualizations be anticipated if event models do not carry such information? The same is true for recognition if modeled closer to human performance. Humans appear to distinguish between typical and atypical events and 'recognize' the typical.

At this point we began our investigations of analogical quantitative motion representations. Analogical representations will be a main theme of the following sections.

In summary, propositional event models are useful for translating visual motion into natural language. The transition from quantitative visual data to qualitative natural language expressions is achieved by applying a basic set of primitive predicates to a basic set of perceptual primitives.

4. Learning object motion

The general goal of learning object motion is to establish trajectory representations with predictive power. Towards this goal it is useful to abstract, to distinguish and to classify trajectories of interest.

Let's assume that a system observes traffic motion in a street scene. The input would be a large number of trajectories being part of many different events and resulting from current and past observations. A desirable output could include a compact description of different event classes like 'overtake'-events, 'turn-off'-events, 'park'-events, 'cross'-events and so on; this representation should allow to predict to recognize, to reason over and to communicate about observed and future events in this domain.

We assume that the paths of all objects are given in spatiotemporal coordinates resulting from low-level perceptual processes. In addition, we

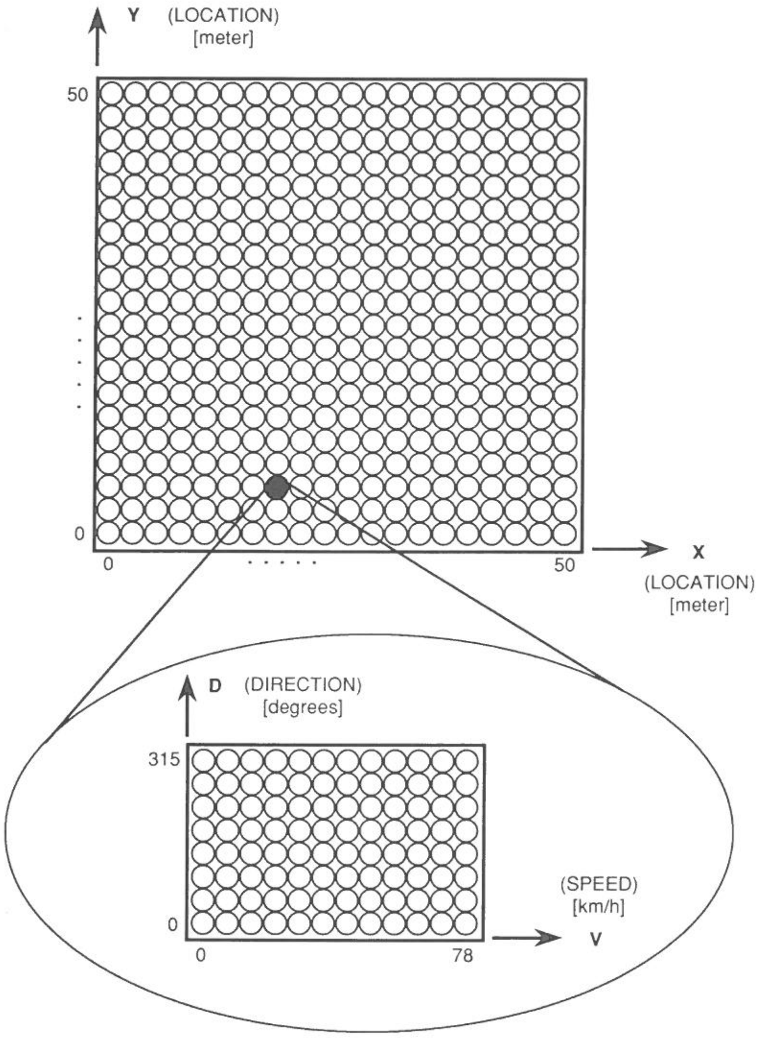


Fig. 7. Spatiotemporal buffer in trajectory accumulation mode. An example with 50 × 50 location cells and for each location cell an array of cells each representing a particular orientation (from 0 to 315 degrees) and a particular speed (from 0 to 78 km/h).

assume that the stationary environment has been analyzed, i.e. all stationary objects are known.

Because we assume no specific a priori knowledge we store elementary information about position and change of position using an exemplar-based approach. It seems natural to consider a representation which is analogical with respect to these dimensions of the modeled world to provide information for learning which is sufficiently rich and also close to perceptual representations. After acquiring sufficient world behavior, relevant information for further abstraction can be made explicit. It will be shown that an analogical representation leads to a natural transition from observed examples to accumulated experience through local operations. We discuss several steps of abstraction from single examples given in a certain concrete geometric environment towards qualitative and propositional event descriptions independent of a particular geometry:

- Accumulation of examples (Section 4.1),
- Generalizations and computation of prototypes (Section 4.2),
- Generic models (Section 4.3),
- Steps towards propositional descriptions of generic models (Section 4.4).

Our approach differs from other exemplar-based models (see e.g. [5,16,40]), because they are exploiting propositional representations. Hence, processes working on the representation look different. It is interesting to note that exemplar-based representations are important for several aspects of human concept formation (see [38]).

4.1. The accumulation of trajectories

We use a spatiotemporal buffer as a representation to accumulate experience about observed trajectories. As a spatiotemporal buffer can support several tasks (see *Fig. 1*) we consider the accumulation of trajectory information as one of several modes of operation. [6]

In this mode the buffer is a four-dimensional accumulator array $C(x, y, d, v)$ covering a certain subfield of the xy-plane. [7] For each xy-pair there are counter cells for all possible velocity vectors, each represented by direction d and speed v (see *Fig. 7*). The vector $S = (x \ y \ d \ v)$ describes the motion state [8] of an object at a given time.

Note that S is composed of quantities which may be perceived by the observer of a visual scene and which only require elementary knowledge about locations and change of locations of identified objects; no further a priori knowledge is needed.

For each observed or remembered object trajectory, a trace of state vectors is registered in the buffer by incrementing the associated counters. As more objects are entered, more cells (possibly the same) are incremented without discriminating between different objects. A trajectory is discretized according to the resolution provided for the different dimensions of the buffer. After observing several examples the counter cell representation exhibits areas of different likelihood without further computation. This is one main advantage of the proposed representation.

The representation can be based on all observed examples or on a more recent set of observed examples by ‘forgetting’ old observations. Forgetting can be performed by slowly decreasing all counter values while new examples are stored.

4.2. Generalizations and prototypes

The output of the trajectory accumulation can be considered a four-dimensional density field with high values indicating experience supported by many observations. Two considerations lead to the next processing steps. First, we cannot assume to obtain examples for every possible situation for which predictions might be needed; hence generalizations from a given set of examples are necessary to cope with situations which differ slightly from the observed examples. And second, there are situations where only the most typical event instances are needed, for example, to support several reasoning tasks (see Section 5); hence an explicit representation of typical trajectories is advantageous. In addition, it can be used for efficient long-term memory.

4.2.1. Generalizations

We introduce the following *generalization* operation: experience represented by a counter cell is propagated to similar trajectories corresponding to its neighbors. Note that similarity between different examples is implicitly given by the Euclidean distance within the analogical representation. The propagation is accomplished by replacing the value of each cell by the weighted average of all neighbors orthogonal to the direction of motion. Cells along the direction of motion contribute according to their positive difference.

In *Fig. 8* the effect of the generalization operation is demonstrated. The left picture shows a

[6] In Mohnhaupt [25] and Mohnhaupt and Neumann [26] we called the representation trajectory accumulation frame (TAF), because the buffer was initially only used for the accumulation of trajectories.

[7] In the domain of street traffic we can restrict ourselves to planar motion.

[8] Our approach differs from the accumulation of physical states used to control dynamical systems (see e.g. [7,24]. One reason is that our representation is organized such that possible prediction (see Section 5) can only be found in the local neighborhood of a certain state. This allows for local prediction and local learning operations.

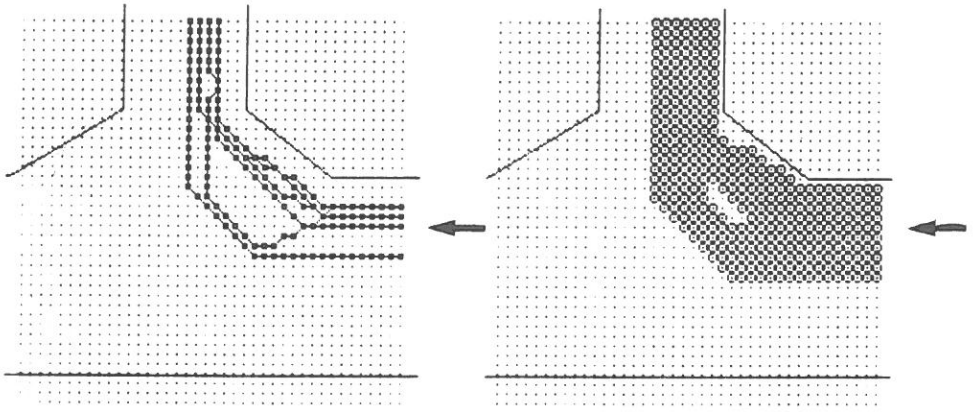


Fig. 8. Buffer filled with several trajectories (left). Information in the same buffer after 'generalization' (right).

buffer filled with eight 'turn-right' examples.[9] Only those cells are activated which are given by the examples. The right picture shows the buffer after applying the generalization operation twice. Information is propagated to similar trajectories in the local neighborhood. This applies also to neighboring velocity values, which cannot be seen in the figure.

4.2.2. *Prototypes*

Traces along density maxima play a special part within the buffer. They form a pattern of typical trajectories in the sense that they outline distinct paths which are maximally supported by experience. We call this pattern a skeleton. Distinct trajectories in a skeleton are called prototypes. The number and the shape of prototypes depend on the set of observed examples. A skeleton might contain a single prominent path or different paths with possibly different degrees of support by the examples. Its shape can correspond to single examples or it can correspond to generalized and averaged information from many different examples. We show how to compute a skeleton and how to code this prototypical information efficiently for the use in subsequent and more abstract stages of processing.

The distribution of the local maxima is a function of the scale at which the density field in the buffer is looked at. At a fine resolution there might be multiple local maxima paths for a 'turn-right' event but on a large scale they might merge into one prominent path. For different tasks different resolutions are needed. Therefore we introduce a *convergence* operation. This operation computes an abstraction of the current buffer by emphasizing trajectories with high support and by suppressing trajectories with lower support. Natural candidates for abstractions are trajectories which are similar with respect to a subset of their properties. Because similarity is reflected by distance within the representation, abstractions can be computed by a local operation. Roughly speaking, to apply the convergence operation, each cell S adds weight to all neighboring cells from where the cell S can be reached. The weight is proportional to neighboring counter values. Therefore cells with higher values get more additional support than cells with lower values. We show the effects of the convergence operation with examples:

Fig. 9 shows two skeletons of the buffer in *Fig. 8*. The left skeleton contains several relative maxima. In the right skeleton the buffer converged into one prominent path.

[9] This figure and the following experimental results show trajectories of moving objects on different intersections from the birds-eye perspective. Trajectories are simulated using a trajectory editor. They are shown as chains of little black squares (see left illustration). Results of local operations are shown as chains of circles (see right illustration). Velocity information which is part of the buffer representation is not visible in the figures.

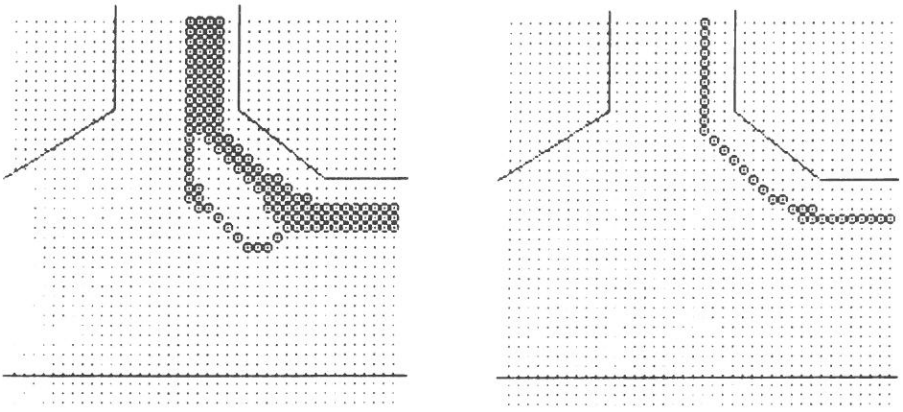


Fig. 9. Skeletons of a buffer after different degrees of convergence.

The left illustrations in *Fig. 10* and in *Fig. 11* show the *xy*-projection of a buffer containing ten trajectories representing different sets of simulated examples. Note that cells with equal *xy*-location but different velocities are distinct in the buffer but cannot be distinguished in the figure. The right illustrations show the skeleton of the buffer after applying the generalization and the convergence operation. The most important paths are made explicit. Note that both skeletons show two conceptual clusters ('turning right' and 'turning left').

The skeletons contain condensed information of the associated buffer: a coarse view of the accumulated experience. It can be the basis for an efficient and abstract description. We use an ex-

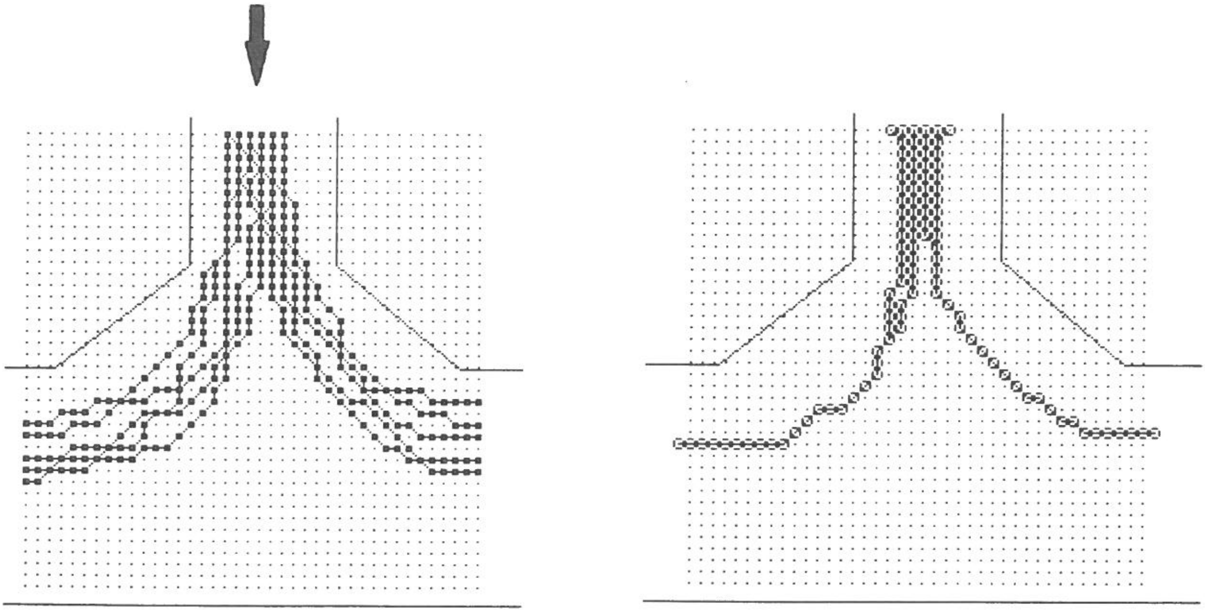


Fig. 10. Buffer filled with ten trajectories (left). Skeleton of the same buffer after 'generalization' and 'convergence' (right).

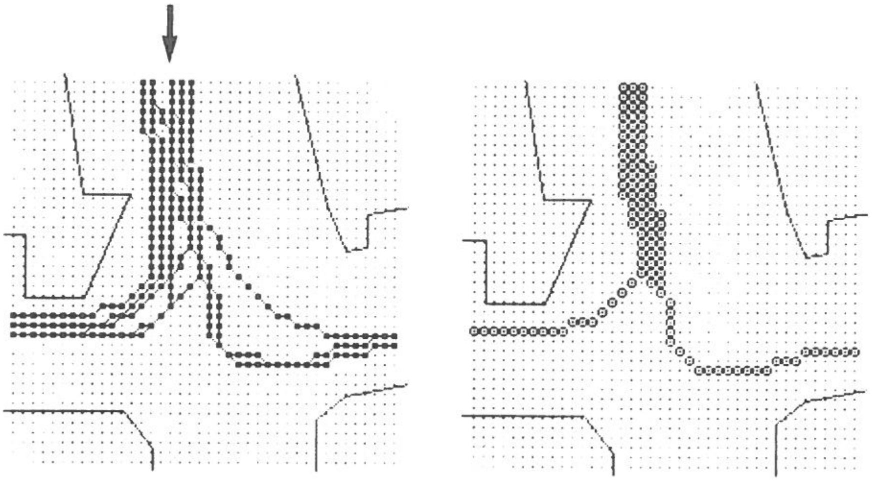


Fig. 11. Buffer filled with another set of observations (left) at a different intersection. Skeleton of the same buffer after 'generalization' and 'convergence' (right).

tended (multidimensional) chain code as an initial description for skeletons. This representation was originally proposed for coding curves in space (see [3]). In space-time it can be used to describe a skeleton in terms of chains of shape elements each of which represents a certain direction and a certain speed at a certain point. Forks and joints are also allowed. The spatiotemporal shape elements of a chain code representation are quite simple, but it is important to note that temporal order is intrinsic within a chain-code representation, because single elements can only be accessed via its temporal predecessors.

Naturally, the chain-code representation can also be used to store individual trajectories independent of the part they play in the accumulator array. This is needed, for instance, if single event examples are temporally far apart. Hence they have to be stored before abstractions can be computed using the buffer.

4.3. Towards generic models

The accumulated experience and resulting skeletons as described in the last section are situation specific in several respects: They are based on examples collected in a particular geometric environment, for example, a particular intersection where 'turn-right' examples could be observed.

A natural next step towards a generic event description is to derive event characteristics which are independent of a particular geometry. Using generic event descriptions, predictions can be made for similar, albeit novel, situations. Also, experience from different situations can be combined.

As an example, we now consider the task of making the experience accumulated at a particular street intersection applicable to another intersection with a different shape. This is an important generalization as we cannot have direct perceptual experience for every possible geometric environment. Note that this generalization step can also be seen as computing an analogy.

The key idea is: First, to enrich the basic trajectory representation by computing perceptual primitives as additional descriptive properties. Second, to select invariant event properties from the set of perceptual primitives for the generic model.

4.3.1. Invariant event properties

Invariant event properties are taken to constitute necessary and sufficient conditions for generic event descriptions. The aim is to derive a particularly convenient set of descriptive primitives given the tasks of interest. The set of invariant event properties should facilitate tasks (be useful for later stages of processing) and should be robustly and efficiently accessible from the data. In addition they should provide a complete event representation with respect to the tasks for which they are needed and they should be relatively independent for compactness.

Primitives need not be independent in a strict sense. Rather independence is defined only in terms of the information explicit (see [19]), i.e. the information accessible with little or no computation according to the primitive operations available in the system. Therefore independence means not easily derivable within the time available. For example, we would call velocity and acceleration of an object to be independent although information about the speed of an object over time contains information about its acceleration. The point is that acceleration is only represented implicitly.

Invariant event properties are in general a subset of the rich repertoire of perceptual primitives exploited by perceptual systems in general (in [44] processes which extract these primitives are called 'visual routines'). Perceptual primitives which show regularities over different examples are candidates for this subset. Naturally the dimensions already available in the buffer *(x y d v)* belong to these candidates. In addition it is useful to include temporal derivatives of these basic physical observables as well as measurements relative to an object of reference. Note that these new dimensions are already implicit in the buffer representation given the availability of reference objects. The following features constitute a useful set of perceptual primitives for characterizing time-varying events in terms of invariants.

(1) location,
(2) orientation,
(3) orientation change,
(4) velocity,
(5) acceleration,
(6) location relative to a reference object (distance),
(7) distance change,
(8) orientation relative to a reference orientation,
(9) orientation change relative to a reference orientation.

The repertoire of perceptual primitives contains some properties which refer only to trajectories (1–5), and some properties which refer to a relation between trajectories and an object of reference (6–9).

It is interesting to note the similarity between these primitives and the primitive events used for bottom-up event recognition in Section 3. Earlier, primitive events were derived by analyzing natural language motion verbs, which are taken to define complex events. These qualitative primitives can be computed from a quantitative scene description comprised of exactly the same quantities as in this repertoire of perceptual primitives. Constant values (like constant velocity or constant motions), restricted values (like 'parallel', 'close to' or 'beside'), comparative values and constant derivatives (like constant acceleration) formed a basic set of primitives. The main concern was to generate a natural language description of time-varying scenes. Therefore a propositional event description based on such primitives is a useful intermediate representation. Because our goal here is to use event descriptions in an image-like context, our representation looks different. Nevertheless, both cases have strong similarities with respect to the set of perceptual primitives and the process of abstraction.

It is an interesting question when to extract the set of invariant properties. In general there are many objects and the number of candidates for invariants grows exponentially with the number of interesting objects. Due to this complexity we believe that invariants cannot be extracted simultaneously, in pace with the changing environment. By having the invariants implicit in a description based on the buffer, this can be left for subsequent reasoning using visualizations and local and parallel processes. This will be touched in the next section; for now we assume invariant properties to be available.

4.3.2. Examples

We describe generic event models collected from observations for 'Turn-off' events and 'overtake' events.

'Turn-off': A detailed description of the buffer loaded with 'turn-off' events has already been given in the previous section. The buffer contains the dimensions *(x, y, d, v)*. As pointed out above, complete information about additional dimensions may not be explicit in the original buffer. After refilling the buffer and under the assumption that the stationary background is known,

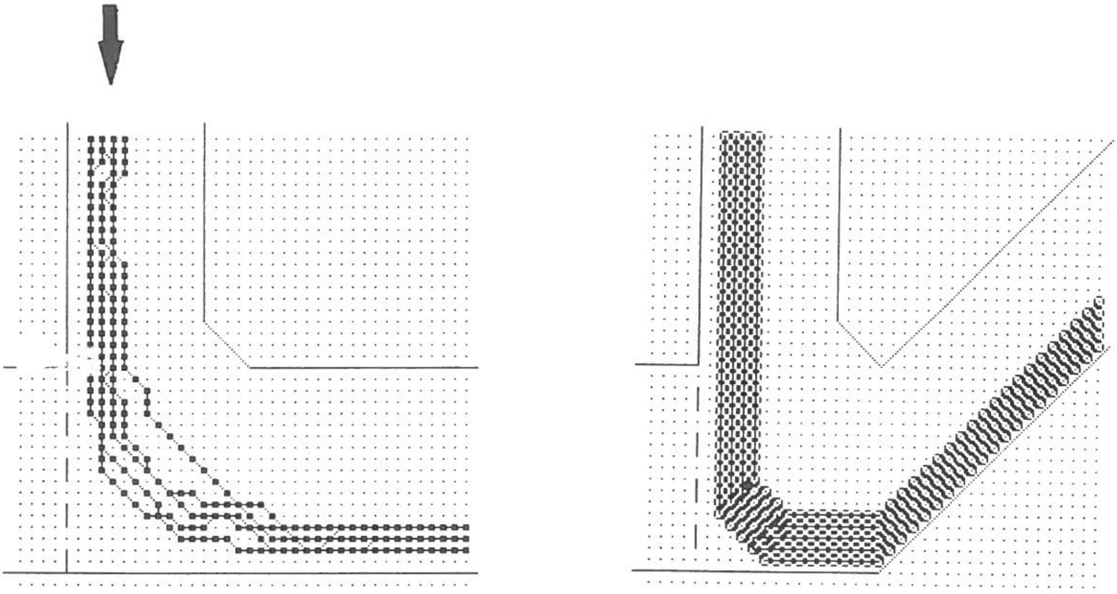


Fig. 12. Five 'turn-off' trajectories on an intersection from which the generic model was derived (left). 'Turn-off' information for the new intersection (right).

invariant dimensions can be made explicit, variant dimensions can be removed.

In the case of 'turn-off', the spatial dimensions *xy* and the orientation *d* of velocity are variant with changing geometry. But the following dimensions capture invariant event properties:

- speed *(v)*,
- relative orientation between car and sidewalk *(ro)*,
- distance between car and sidewalk *(di)*.

In the generic model of a turn-off event, speed information consists roughly of three parts: a constant deceleration at the beginning of the event, constant speed at the intersection and a constant acceleration at the end of the event. The relative orientation between car and sidewalk and the distance between car and sidewalk remains almost constant over the whole intersection. This might change in a more detailed model but it serves as a good first approximation. In order to learn a generic event model, the invariant information has to be identified and determined from a possibly large number of examples.

The following example demonstrates the use of a generic 'turn-off' model. A buffer for 'turn-off' is instantiated for a new intersection using information derived from another intersection which differs in shape.

The left illustration in *Fig. 12* shows five 'turn-off' trajectories on an intersection. Based on these event instances the invariant perceptual primitives were computed to build up a generic model. The original dimensions like *xy*-locations are only stored implicitly in the generic model. In order to apply the information to the particular geometry of the intersection in the right figure they have to be recomputed. The right illustration of *Fig. 12* shows information instantiated from the generic model derived from the street shape in the left illustration.

Information collected in a particular environment in space-time has been transformed to be applicable in a different environment by conserving event dependent information and abstracting away irrelevant information. This figure can also be interpreted as showing the computed visualiza-

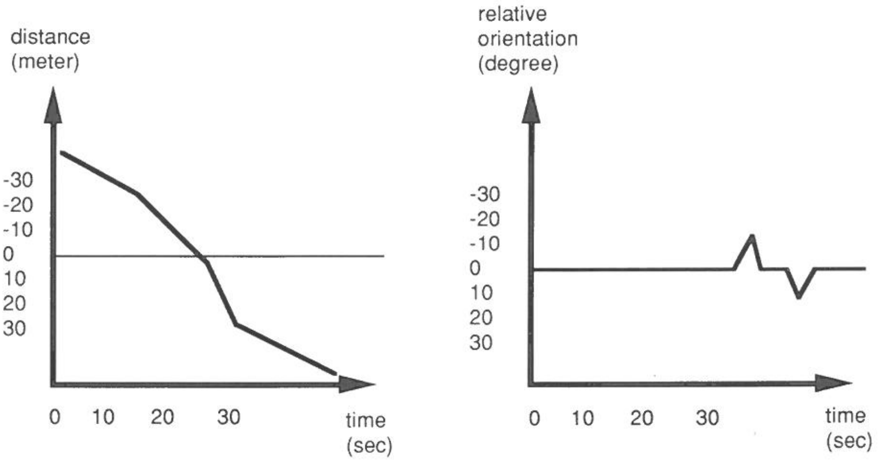


Fig. 13. Generic Model for 'overtake'.

tion of 'turning-right' events on intersection 2 in analogy to observed 'turning-right' events on intersection 1.

'Overtake': Let's turn to another example. After observing several 'overtake' events between two cars on different locations of a street scene, a generic model can be based on:

- the distance between the two cars *(di)*,
- their relative orientation *(ro)*, and
- their relative speed *(rs)*.

The observed trajectories follow a typical path through this 3-dimensional space. For example, distance and relative orientation over time might look as shown in *Fig. 13*.

The generic model can be viewed as a trace through a more-dimensional space like the skeleton introduced above. Hence, the information can in principle be stored by coding segments of the path using the chain code as described for skeletons.

Following Palmer's definition [32] the generic model is analogical with respect to its dimensions. To bridge the gap towards abstract language ori-

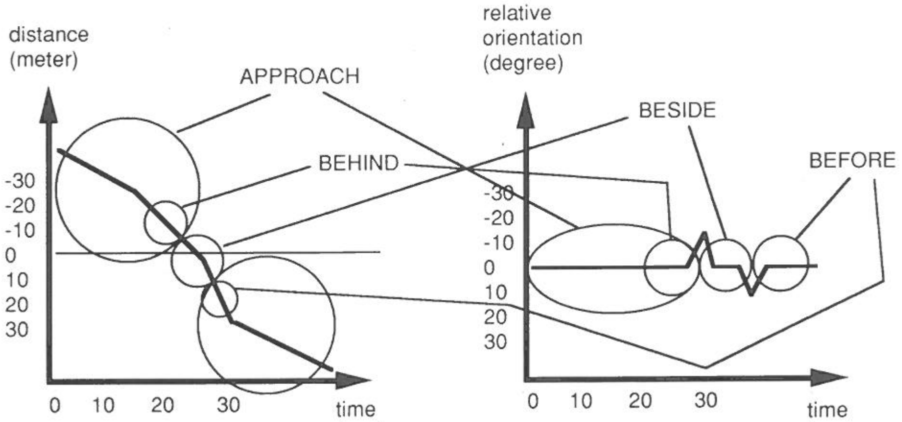


Fig. 14. Predication over a generic model for 'overtake'.

ented event models as used in Section 3, a propositional description of generic models, based on qualitative predicates, will be derived in the following subsection.

4.4. Towards propositional generic models

Some tasks related to object motion require abstract propositional event descriptions, for example, for verbal communication (see Section 3). A natural next step towards further abstractions is to describe the generic event model by some of its meaningful segments or points. This leads to more compact and qualitative descriptions which can be used for a subsequent predication.

We turn to the 'overtake' example from the previous subsection. We derive here an intuitive predication. The values which characterize the distance between the two cars and their relative orientation can be segmented into the following intervals.

The segments correspond roughly to 'approach', 'behind', 'beside', 'in-front-of' and 'recede'. Of course, it is a complex problem to define the right set of primitives for a propositional characterization of an event. This can not only be based on the event to be described but has to be seen in comparison with other events. We have not yet developed an algorithmic solution, but we can show that using the generic event model the appropriate dimensions for a subsequent predication are made explicit and therefore facilitate this task.

5. Spatiotemporal reasoning

In this section we discuss several relevant spatiotemporal reasoning tasks. The tasks are all based on visualizations within the analogical spatiotemporal buffer and local processes working on it. We view this kind of reasoning as a useful alternative for propositional and logic-based reasoning about the environment. The restricted analogical representation allows to reduce complexity for several tasks because spatiotemporal relations are easily derivable and the representation is complete with respect to space and time.

First, we describe the instantiation of a spatiotemporal buffer from long-term memory (5.1). Second, we discuss the visualization of single and multiple trajectories (5.2). Third, we show how to compute spatiotemporal relations using the buffer and local spreading activation processes (5.3), and fourth, we discuss how event models can be constrained by current perceptual data. This will allow for reasoning in situations which slightly differ from the accumulated experience. For example, in street traffic scenes with obstacles (5.4).

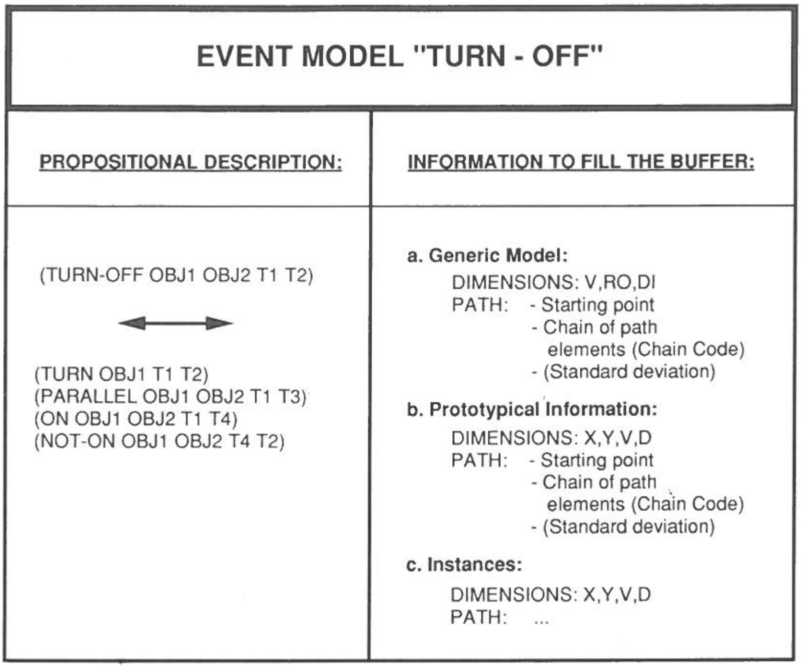
5.1. Instantiation of a spatiotemporal buffer

The spatiotemporal buffer is instantiated on demand from long-term memory. Depending on the system's current knowledge the long-term memory contains:

- Information about instances from particular geometric environments; no generalizations and abstractions have been computed so far. The examples can be filled into the buffer without further computation because they are represented in buffer dimensions.
- Information about prototypical paths (skeletons) for a given geometric environment. The skeleton can be filled in by loading the chain code description. The dimensions of the skeleton correspond to the dimensions of the buffer.
- Information about a generic event model as described in the last section. For example in case of 'turn-off', the generic description contains qualitative descriptions of the dimensions 'distance to the sidewalk', 'relative orientation to the sidewalk' and 'speed'. This generic information has to be adapted to a new instance (the actual geometric environment) by reintroducing the dimensions *(x, y, d)*, which depend now on the new location. Each point of the new intersection has a certain distance to the sidewalk and a certain orientation relative to the sidewalk. Its value is set according to the activity in the generic description for this combination of coordinates.

In addition, the buffer can be filled with actual data from low-level vision processes, for example, information about the current static scene parts, obstacles on the street, etc. The buffer is a shared representation between perceptual and cognitive processes.

In *Fig. 15* an example of a 'turn-off' event model in long-term memory is shown. As described above, the systems knowledge can be in three possible states. Note that they do not exclude each other. For example, the system might



EVENT MODEL "TURN - OFF"

PROPOSITIONAL DESCRIPTION:

(TURN-OFF OBJ1 OBJ2 T1 T2)

(TURN OBJ1 T1 T2)
(PARALLEL OBJ1 OBJ2 T1 T3)
(ON OBJ1 OBJ2 T1 T4)
(NOT-ON OBJ1 OBJ2 T4 T2)

INFORMATION TO FILL THE BUFFER:

a. Generic Model:
DIMENSIONS: V,RO,DI
PATH: - Starting point
- Chain of path elements (Chain Code)
- (Standard deviation)

b. Prototypical Information:
DIMENSIONS: X,Y,V,D
PATH: - Starting point
- Chain of path elements (Chain Code)
- (Standard deviation)

c. Instances:
DIMENSIONS: X,Y,V,D
PATH: ...

Fig. 15. 'Turn-off' event model in long-term memory.

have a generic model for 'turn-off' as well as a specific model for a particular intersection at the same time.

5.2. Visualization of trajectories

Let us consider now the visualization of trajectories using an instantiated spatiotemporal buffer. We can think of different situations where this task is useful. In general it can provide predictions about a scene based on an event model. For example, given starting points of observed objects and assumptions about the event to be seen, predictions allow for a significantly constrained scene analysis (see [27,29]). In addition, visualization of prototypical trajectories allows for spatiotemporal reasoning about the collision of objects. For example, given starting points of several objects, a likely accident can be predicted. As described in Section 3, visualizations are also advantageous for several communication tasks; for example, using visualizations in natural language understanding one can check the propositional content for consistency and completeness (see also [1,14,45]). In the next section it will be shown how visualizations can be used for computing perceptual primitives.

We now describe the visualization algorithm. Given a starting cell the obvious operation to visualize a typical trajectory is to look for the maximal counter in the four-dimensional vicinity. Note that not all xy-neighbors are eligible if the velocity direction is restricted to vary smoothly. *Fig. 16* demonstrates possible predictions. The successor neighborhood represents possible successor cells of cell $(n\ m\ v\ d)$. In the figure, the orientation dimension d is represented by the orientation of arrows, and speed v is represented by the length of the arrows. When visualizing the most typical trajectory, the cell with the highest activation value is chosen from the successor neighborhood. In the case of visualizing a bundle of likely trajectories, all elements above a certain activation threshold are chosen (see examples). The predecessor neighborhood is also important, it will be discussed in Section 5.4.

The left pictures in *Fig. 17* illustrate the visualization of single trajectories given different start-

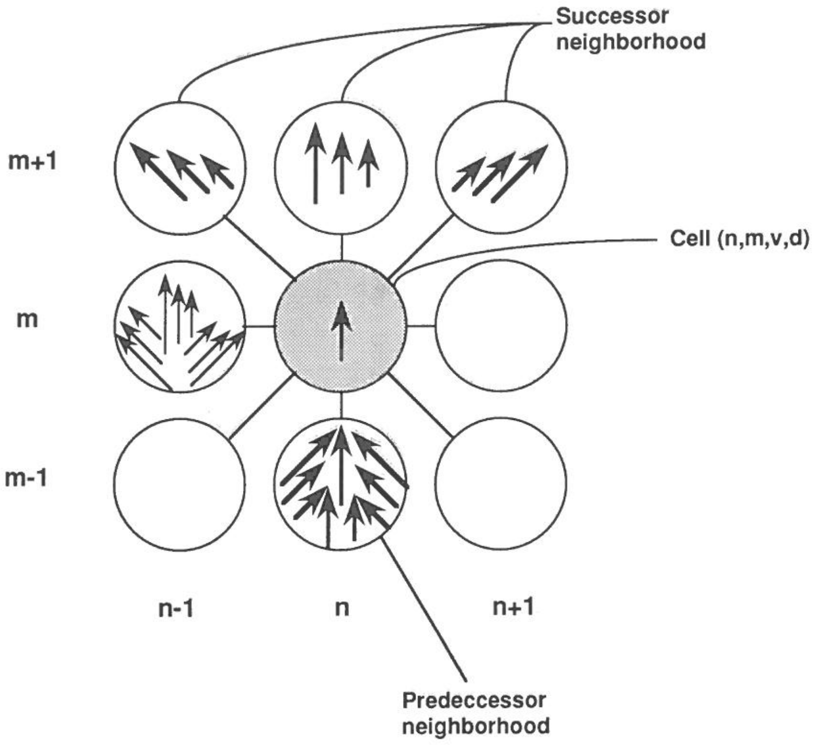


Fig. 16. Visualization algorithm.

ing points (solid). The right illustrations show bundles of possible trajectories which are most likely given the starting points. A bundle can be used, for example, as search area for top-down controlled image sequence analysis. The upper illustrations are based on the simulated examples in *Fig. 10* after generalization and convergence operations. The lower illustrations are based on the examples in *Fig. 11*. Other examples of the prediction algorithm can be found in Mohnhaupt [25] and Mohnhaupt and Neumann [26].

It is interesting to note that a prediction is physically plausible because it relies on observations of physical behavior and plausible generalizations thereof. Hence, the constraint of physical plausibility is internalized in the representation; no further computation is necessary.

The same trajectory visualization mode is used in multiple object situations. Given starting points of objects, their temporal relations, and an event model of their future behavior (one might be a 'turn-off', the other might be an 'overtake' event), collision of object paths can be predicted. By allowing only one object per *xy* position, physical plausibility is maintained. In addition, trajectory prediction can be used to support event recognition as discussed in Section 3. The trajectory of a recognized event is compared to the typicality distribution within the buffer to decide how typical or atypical the event is.

5.3. Computing spatiotemporal relations

In Section 4 we described the importance of perceptual primitives for the extraction of invariant event properties. Perceptual primitives express spatiotemporal relations. They are of general interest for important tasks in any vision system including path-planning, obstacle avoidance, and so on. Therefore, robust and fast extraction of spatiotemporal relations is crucial for intelligent systems.

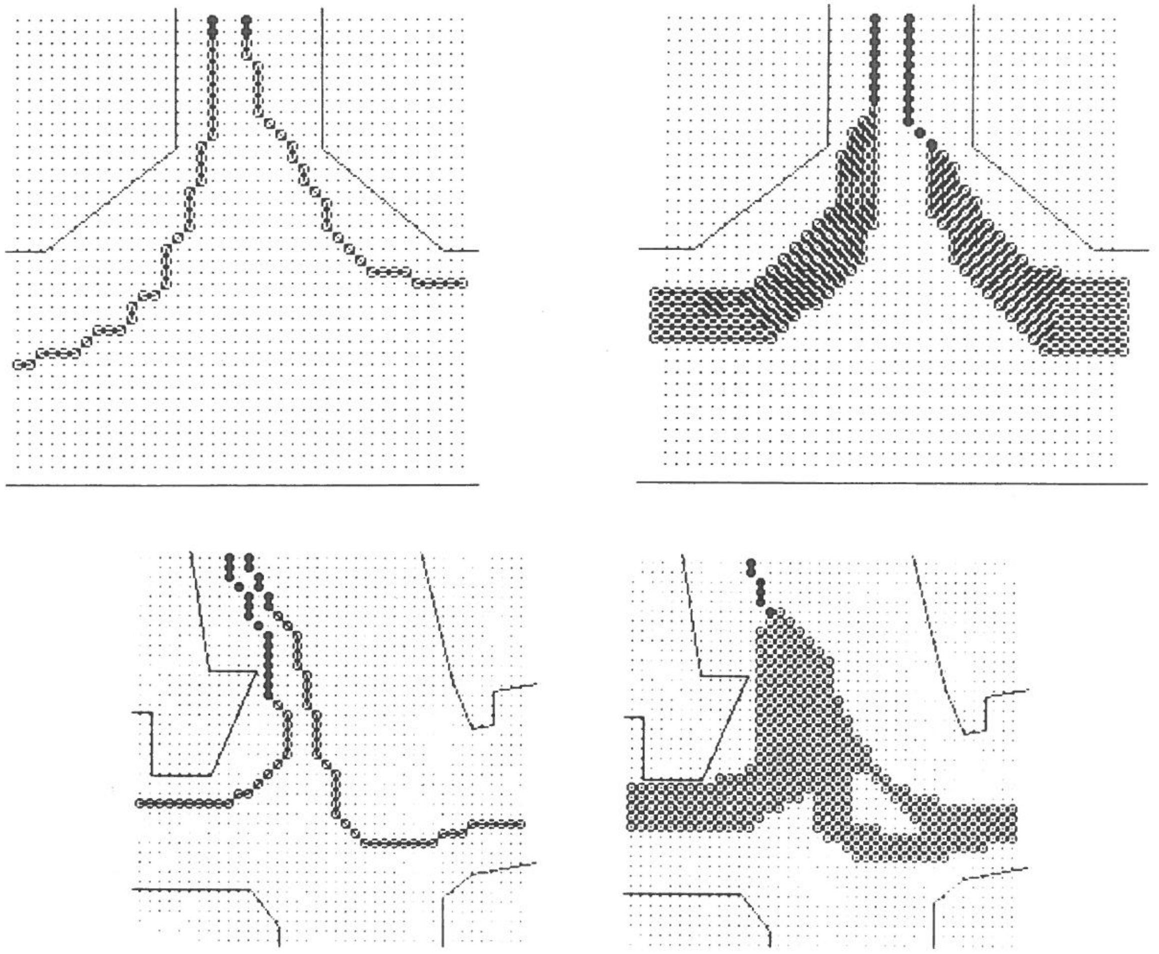


Fig. 17. Single visualized trajectories given different starting points (left). Bundles of visualized trajectories (right).

The spatiotemporal buffer representation allows for the measurement of spatiotemporal relations by simple spreading activation processes. The reason is that spatiotemporal relations are explicitly represented within the analogical spatiotemporal buffer. We demonstrate the use of spreading activation processes with an example.

The task in the example is to measure the perceptual primitive 'distance' between object 1 and object 2 as shown in *Fig. 18*. Object 1 (which covers a certain position in *xy*) sends a pulse of activation to its spatially adjacent cells. Upon this the neighbors send a pulse to their neighbors with a certain decay of activation and so on. Each cell which belongs to an object 'remembers' the maximum level of activation of all incoming signals. Hence, lines of equal activation represent lines of equal distance. After the activation process is

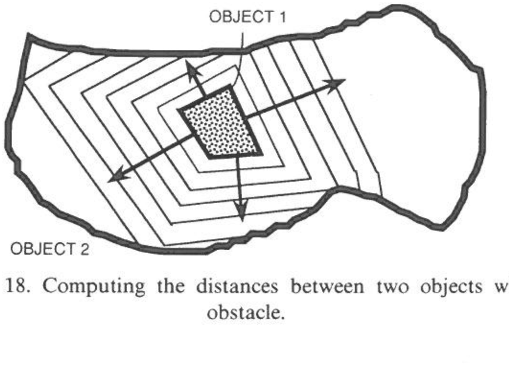


Fig. 18. Computing the distances between two objects with obstacle.

finished, each cell of object 2 keeps the shortest distance to object 1 coded as activation value. The spreading activation process can also be modeled as reaction diffusion rule (see [41]).

While there are efficient ways to compute spatial relations, we do not assume that all perceptual primitives can be extracted in real-time in pace with the incoming information. The reason is the possibly very large number of potentially interesting relations and the need to consider several examples in order to find relevant primitives. Therefore the buffer is initially used to build up basic information about location and velocity in an intermediate representation (see *Fig. 15*). A subsequent visualization process using the instantiated spatiotemporal buffer is then explored to compute perceptual primitives. Hence, initial storing and subsequent processing like learning and reasoning can be separated. This is a useful computational feature of the proposed representations. It can also be viewed as reinterpreting the initial observations.

There is a dispute in the 'imagery' literature about possible reinterpretations of mental images. Finke et al. [10] present evidence for the possibility of reinterpreting mental images and Reisberg and Chambers [35] argue against it using different empirical results. From a computational point of view a reinterpretation might be necessary and useful, for example, for computing additional perceptual primitives.

In principle, relevant spatiotemporal relations could also be measured on a qualitative propositional level using logic-based deduction. Unfortunately such representation and mode of processing leads to undesirable complex operations. In *Fig. 19* a more complex distance measuring situation is shown including an obstacle. The task is to determine the shortest-path distance from object 1 to object 2. By using the buffer and spreading activation processes the complexity remains the same in spite of the obstacle. But consider the case of using propositional descriptions of all objects in terms of coordinate pairs. Euclidean distances for obstacle-free situations could be easily computed. With obstacles, however, a much more complex shortest-path computation would have to be carried out.

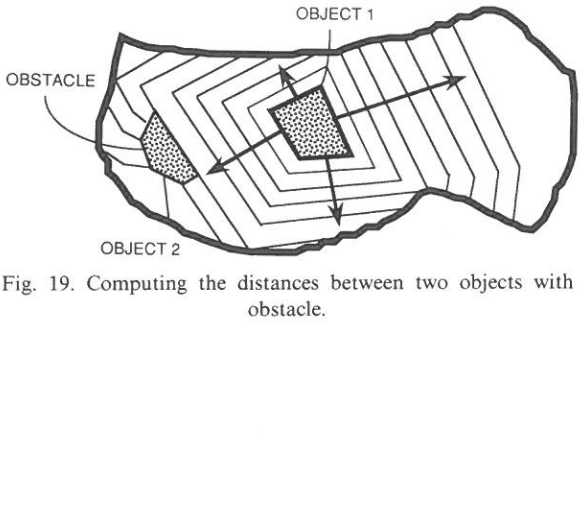


Fig. 19. Computing the distances between two objects with obstacle.

5.4. Constraints on event models

Now we want to show how event models represented in the spatiotemporal buffer can be adjusted to cope with slightly different situations, e.g. situations with obstacles which were absent when the experience was accumulated. We show that the representation is flexible enough to allow meaningful predictions in these situations. We also use this example to demonstrate the integration of cognitive representations (event models) with perceptual data (obstacles in a given situation) in the spatiotemporal buffer.

An obstacle is a subspace of the 4-dimensional buffer where activities are not allowed, for example, due to a parking car or due to a forbidden range of velocities (in case of snow on the street). We introduce an obstacle into the buffer by setting the counter values of the appropriate subspace to zero. Then we propagate this information through the buffer by using a local inhibition operation.

Inhibition: We inhibit a cell S by setting its counter value to zero if:

- All the counter cells which can be reached from S are equal to zero (successor neighborhood, see *Fig. 16*),
- or all the counter cells from which S can be reached are equal to zero (predecessor neighborhood).

This operation is performed repeatedly for all cells until no more changes occur. Thus, one is sure that all cells which have no active predecessor or no active successor are set to zero. We demonstrate the inhibition operation with examples.

The left illustration of *Fig. 20* shows a buffer filled with information about 'turn-off' events after introducing two obstacles. The effects of the inhibition operation are visible in the right illustration. All trajectories which would pass through the

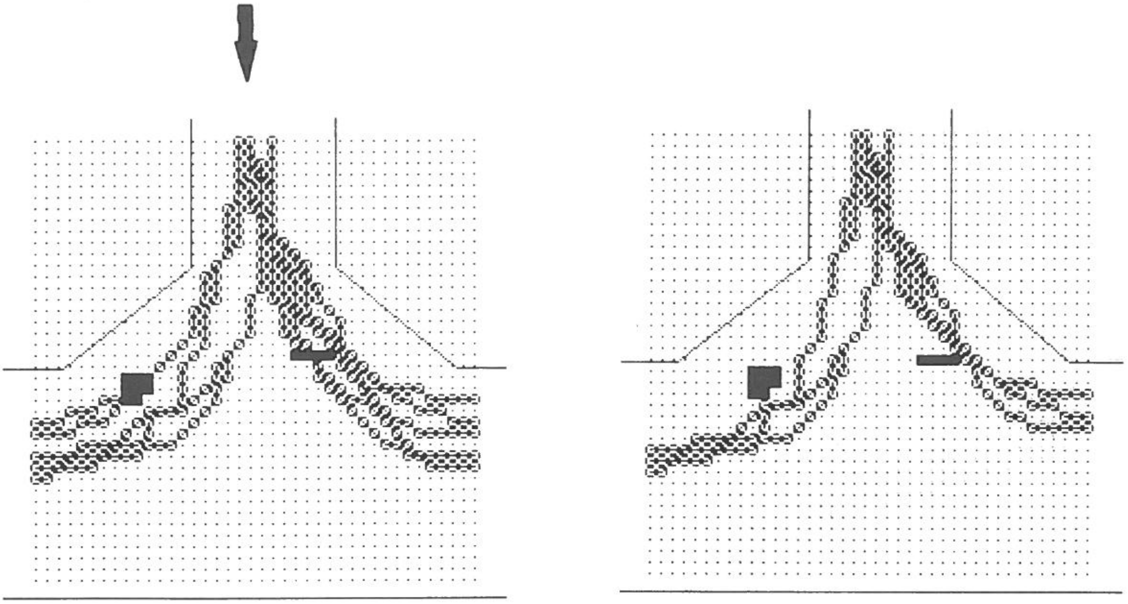


Fig. 20. Skeleton of a buffer containing 'turn-off' examples with obstacles (left). The same skeleton after inhibition (right).

obstacle are inhibited. Hence the representation is adapted to the constrained situation.

The left illustration of *Fig. 21* shows a skeleton within a buffer after introducing an obstacle that does not allow for any of the prototypical paths. The effects of the inhibition operation are visible in the right illustration. The computed trajectories avoid the obstacle. The limits of this kind of

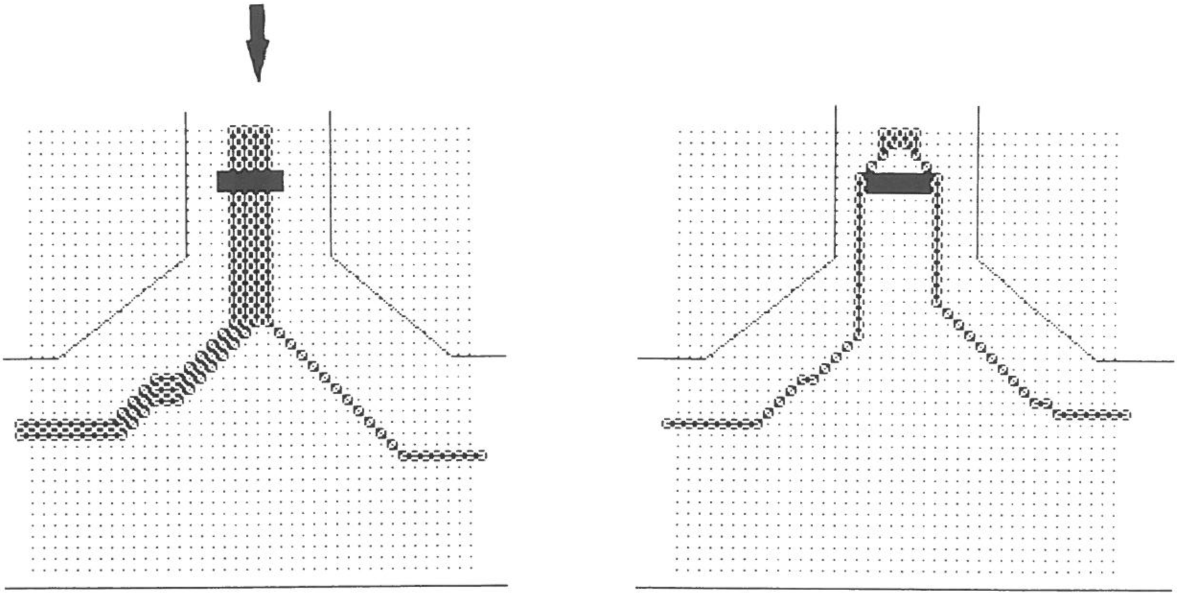


Fig. 21. Another skeleton of a buffer with an obstacle (left). The skeleton after inhibition (right).

obstacle avoidance are set through the observed and generalized examples. If none of the stored paths (obtained from observations and generalizations) allows for obstacle avoidance, a meaningful prediction is impossible.

This example shows the effects of inhibition for predictions given particular starting points. The predictions are based on the accumulated examples in *Fig. 8*. After inhibition a previously unplausible prediction (left) is appropriately adapted to the obstacle.

6. Representation issues

Choosing an adequate representational scheme and mode of computation is certainly a key problem for intelligent systems. There are few sound principles to favor one representation over another. Despite this it is instructive to view the current approach from this perspective.

Following Palmer [32], who is concerned with a computational metatheory for representation, a representational system consists of two worlds, namely, a represented world and a representing world. The objects of interest in the represented world are mapped onto the representing primitives. The function of the representation is to preserve information (or structure) from the represented world. Thus, constraints are imposed on the representing world in a way that reflects relations between their associated objects in the world. We adopt several points of Palmer's theory and describe differences and extensions resulting from our work. Other relevant discussions about the usefulness and applications of analogical/quantitative vs. propositional/qualitative representations can also be found in e.g. Sloman [37], Sober [35], Levesque [19], or Janlert [15].

At least two steps are necessary to develop a representational theory on which the choice of adequate representations can be based. First, one has to identify general constraints on representations as well as constraints resulting from the domain of interest, and second one has to identify general features of representations which allow to overcome the limitations imposed by these constraints.

6.1. Constraints on the representation

Learning: Our work shows that viewing learning as a fundamental concern puts extra constraints on the representation and associated processes. The exemplar-based approach using the spatio-temporal buffer leads to a natural transition from observations to accumulated and possibly prototypical experience. A distance measure for abstractions and generalizations is given by the representation without further computation. Moreover, physical plausibility of the models is intrinsic because they are based on physically plausible observations.

Temporal constraints: The temporal order of the world is a general constraint which should be

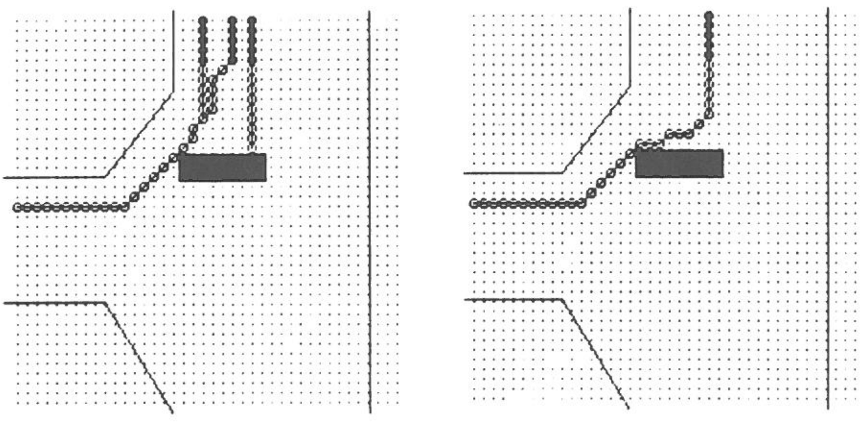


Fig. 22. Predictions before inhibition (left). Prediction of a previously blocked trajectory after inhibition (right).

reflected by an adequate representation. In addition, in the area of object motion there are certain intrinsic temporal constraints of the domain. For example, in case of moving obstacles, the time available to prevent possible collisions is limited.

Frame problem: Shoham [36] investigates the so-called frame problem. In discussing how people solve a prediction task concerned with a billiard game he concludes: "*The inevitable answer seems to be that they 'visualize' the problem, identify a solution in some mysterious ('analog') way, and only then validate the solution through physics*"

We see our work as an attempt to clarify this mysterious analog way. Physical plausibility is maintained at low cost by relying on concrete observations and abstractions thereof as well as by representing basic physical knowledge (e.g. one object per point in space) intrinsically. In addition, the spatiotemporal buffer representation guarantees that relevant spatiotemporal relations change automatically through reasoning by visualizations and local processes. Consistency of a description of spatiotemporal occurrences at a certain point in time can be easier maintained within the quantitative analogical representation because it is constrained in the right way. The propositional qualitative representation is more general and allows to represent almost everything, but for the price of extra costs to maintain constraints implied by a particular domain.

6.2. Features of the representation

Intrinsic and extrinsic constraints: Palmer [32] calls a representation intrinsic whenever "*the representing relation has the same inherent constraints as its represented relation*". Thus the structure of the represented world is preserved through the choice of a primitive vocabulary, accompanied by its own intrinsic constraints. [10] One major reason for the past use of analogical representations has been their intrinsic constraints on spatial relations such as distances, areas, relative locations, and orientations. Our work suggests that similar arguments apply in space-time. In addition, important physical constraints can be maintained using the analogical buffer.

Explicit and implicit information: The terms explicit and implicit refer to the mode of processing, where explicit refers to accessibility with little or no cost; i.e., the information is extracted using only primitive operations. [11] In general one must reconcile memory costs, task requirements (temporal priorities), processing capacities, and the time required during knowledge acquisition. As more information becomes explicit, redundancy results and efficiency of memory decreases. Exactly what knowledge is made explicit depends on the frequency with which such information is required, and the priority of the tasks for which it is required.

Within the spatio-temporal buffer relevant dimensions for describing object motions are either explicit (like xy), or they can be made explicit through the use of visualizations and local processes.

Informational and computational equivalence: This issue is related to the previous one. Two representations are informationally equivalent if the information derivable is equivalent; they are computationally equivalent if the same information is derivable with the same costs. Informational equivalence is only of theoretical interest. If some piece of information is only implicit and the task it is needed for requires an immediate reaction, then the information is not derivable in practical terms. For example, a human being should have a fast access to the relevant models if a tiger is approaching. In case the information what a tiger looks like is only implicit, its derivability is only of academic and postmortem interest. Many aspects of spatiotemporal relations are only implicitly stored in our long-term qualitative event representation. Therefore we introduced a restricted analogical representation specialized for several tasks and instantiated by demand. The two representations are informationally equivalent, but in the latter relevant relations are explicit and therefore more easily accessible.

[10] Palmer suggested that the intrinsic-extrinsic distinction accounts for the analogical-propositional distinction (imagery debate) (see e.g. [4,17,33]).

[11] When discussing the information content of a representation, in contrast to Palmer [32], we feel that it is useful to specify the derivability of information as a function of cost (in time and space).

A common representation for perceptual and cognitive processes: Perception and cognitive processes are often investigated separately in the literature. Analogical quantitative representations are still questioned in the reasoning community, but are well established in the vision literature (see e.g. the 2 1/2-D Sketch representation in Marr [21]). Our belief is that cognition merges smoothly with perception, thereby implicating shared representational structures and processing capabilities (e.g., see Finke [9]).

A shared representation also makes the basic perceptual apparatus available to the cognitive system with which to solve problems. For it is often expedient to place cognitive tasks within a representational structure that is well-suited to, and constrained by our perceptual model of the world. The study of recognition, prediction, and learning of object motion within the spatiotemporal buffer is simply one step toward a possible understanding of this 'perceptual-cognitive junction'. Consequently, the spatiotemporal buffer allows for the integration of bottom-up and top-down information within one representation. This was shown in the last section and in the area of top-down controlled image analysis, where model-based predictions about ongoing object motion computed by the cognitive system have been used to focus and enhance low-level motion analysis (see [27,29]).

Local and parallel style of processing: The style of processing within a representation has severe consequences for the performance in certain tasks. From this perspective the transition from the qualitative/propositional event models to the quantitative/analogical spatiotemporal buffer can be seen as '*organizing the representation such that local and parallel processes become useful*'. This is a general rule for the design of adequate representations.

The processing within the buffer can also be viewed as subsymbolic following the perspective given by Steels [41]. In his work tasks like reasoning about the behavior of liquids are performed on a subsymbolic layer. In addition, he presents evidence that even very abstract problems (like the 8-puzzle) can be solved within such a framework.

7. Summary

We have developed a framework for understanding object motion. We presented a hybrid system of representations and associated processes to solve relevant tasks related to object motion. The two main representations are:

(1) A qualitative propositional representation including logic-based reasoning processes for event recognition, verbalization, and long-term memory.

(2) An analogical quantitative spatiotemporal buffer including local processes for several learning and spatiotemporal reasoning tasks. This representation is initialized from long-term memory on demand.

Event recognition is performed using hierarchical matching. This takes propositional primitives describing the current scene as input and generates instantiated events as output. The propositional primitives are qualitative predicates on perceptual primitives.

Several learning tasks have been investigated. First, we accumulate examples using the spatiotemporal buffer. Second, we use local operations for computing generalizations and abstractions. This leads to skeletons containing condensed information about the buffer which can be used for efficient storage. We have shown how to exploit a rich set of perceptual primitives to compute invariant event dimensions and to obtain generic event models which are independent of a particular scene geometry. Finally we have sketched how a predication of generic models can lead to the propositional event models used for recognition.

In addition, we have discussed the use of accumulated experience for spatiotemporal reasoning tasks concerned with concrete visual data. These data are either given from perceptual processes or instantiated from long-term memory. Reasoning is performed by visualizations and local processes. We have described the prediction of single and multiple typical trajectories and the computation of perceptual primitives through spreading activation processes. Predicted trajectories can be adapted to constrained situation, e.g. obstacles in traffic scenes. Predictions can also be used to control visual processes, to reason about collision of objects and to decide how typical or atypical a recognized trajectory is.

The hybrid representation has been motivated by general constraints and by constraints ın the domain of object motion: learning as a fundamental concern, temporal constraints, and the frame problem. This leads us to develop an analogical representation which is specialized for several problems of this domain. It features intrinsic physical constraints and explicit representation of spatiotemporal information. In addition, it serves as a common representation for perceptual and cognitive processes, which supports the use of a local and parallel style of processing. A propositional representation at this level could be informationally equivalent and more general, but for the price of extra cost to maintain the constraints of the domain of object motion.

Acknowledgements

We thank Hans-Joachim Novak for his collaboration on event recognition and verbalization, and Klaus Strobel for the 3-D graphics package. Also, we thank David Fleet for many interesting discussion and critical comments on earlier drafts this paper.

References

[1] G. Adorni and M. Di Manzo, Top-down approaches to scene interpretation, *Proc. CIL 83*, Barcelona, Spain (June 1983).

[2] E. Andre, G. Bosch, G. Herzog and T. Rist, Characterizing trajectories of moving objects using natural language path descriptions, *Proc. 7th ECAI* (1986).

[3] D.H. Ballard and C.M. Brown, *Computer Vision* (Prentice Hall, Englewood Cliffs, NJ, 1982).

[4] Ned Block, ed., Imagery. (MIT Press, Cambridge, MA, 1981).

[5] G. Bradshaw, Learning about speech sounds: the NEXUS project. *Proc. 4th Int. Workshop on Machine Learning,* Irvine (1987) 1–11.

[6] R.A. Brooks, Planning collision free motion for pick and place operations, in: M. Brady and R. Paul, eds., *Robotics Research* (MIT Press, Cambridge, MA, 1984).

[7] M.E. Connell and P.E. Utgoff, Learning to control a dynamic physical system, *Proc. Nat. Conf. on Art. Intell. AAAI-6* (1987) 456–460.

[8] M. Denis, J. Engelkamp and J.T.E. Richardson, *Cognitive and Neuropsychological Approaches to Mental Imagery* eds., (Martinus Nijhoff, 1988).

[9] R.A. Finke, Theories relating mental imagery to perception, *Psychological Bulletin* 98 (1985) 236–259.

[10] R.A. Finke, S. Pinker and M.J. Farah, Reinterpreting visual patterns in mental imagery, *Cognitive Science* 13 (1989) 51–78.

[11] K.D. Forbus. Qualitative reasoning about space and motion, in: D. Gentner and A.L. Stevens, eds., *Mental Models* (Lawrence Erlbaum, Hillsdale, NJ, 1983) 53–74.

[12] B.V. Funt, Problem solving with diagrammatic representations, *Atificial Intelligence* 13 (1980) 201–230.

[13] Francesco Gardin and Bernhard Meltzer, Analogical representation of naive physics, *Artificial Intelligence* 38 (1989) 139–159.

[14] Ch. Habel, Propositional and depictorial representation of spatial knowledge: The case of path-concepts, Technical Report, University of Hamburg, FBI-HH-M-171/89 (1989).

[15] L.-E. Janlert, Pictorial knowledge representation, *Proc. 8th ECAI*, (1988) 149–151.

[16] D. Kibler and D.W. Aha, Learning representative exemplars of concepts: An initial case study, *4th Int. Workshop on Machine Learning,* Irvine (1987) 24–30.

[17] S.M. Kosslyn, *Image and Mind*. (Harvard University Press, Boston, MA, 1980).

[18] J.H. Larkin and H.A. Simon, Why a diagram is (sometimes) worth ten thousand words, *Cognitive Science* 11 (1987) 65–99.

[19] H.J. Levesque, Making believers out of computers, *Artificial Intelligence* 30 (1986) 81–108.

[20] R.K. Lindsay, Images and inference, *Cognition* 29 (1988) 229–250.

[21] D. Marr, *Vision* (W.H. Freeman, San Francisco, CA, 1982).

[22] M. Marschark, The functional role of imagery in cognition, in: M. Denis, J. Engelkamp and T.E. Richardson (1988) 405–417.

[23] R.S. Michalski, J.G. Carbonell and T.M. Mitchell, *Machine Learning I* (Tioga Publishing Company, 1983).

[24] D. Michie and R.A. Chambers, Boxes: An experiment in adaptive control, in: E. Dale and D. Michie, eds., *Machine Intelligence II* (Oliver and Boyd 1968).

[25] M. Mohnhaupt, On modelling events with an analogical representation, *Proc. German Workshop on Artificial Intelligence GWAI-11* (1987) 31–40.

[26] M. Mohnhaupt and B. Neumann, Szenehafte Modelle für zeitabhängige Ereignisse, Technical Report, FBI-B-127, Fachbereich Informatik, University of Hamburg (February 1987).

[27] Michael Mohnhaupt and David Fleet, Raum-zeitliche Filter für eien top-down Steuerung der Bewegungsanlage, *Proc. German Workshop on Artificial Intelligence GWAI-12* (1988) 296–305.

[28] M. Mohnhaupt and B. Neumann, Some aspects of learning and reorganisation in an analogical representation, in: K. Morik, ed., *Proc. International Workshop on Knowledge Representation and Knowledge (re)Organisation in Machine Learning* (Springer Verlag, 1989) 50–64.

[29] M. Mohnhaupt and B. Neumann, On the use of motion concepts for top-down control in traffic scenes, (submitted for publication).

[30] B. Neumann, Natural language description of time-varying scenes, in: D.L. Waltz, ed., *Semantic Structures* (Erlbaum, Hillsdale, NY, 1989) 167–207.

[31] B. Neumann and H.J. Novak, Event models for recognition and natural-language description of events in real-world image sequences, *Proc. Int. Joint Conf. on Art. Intell, IJCAI-8* (1983) 724–726.

[32] S.P. Palmer, Fundamental aspects of cognitive representation, in: E. Rosch and B.B. Lloyd, eds, *Cognition and Categorisation* (Erlbaum, Hillsdale, NY, 1978).

[33] Z.W. Pylyshyn, *Computation and Cognition* (MIT Press, Cambridge, MA, 1984).

[34] K. Rehkämper, Mentale Bilder – Analoge Repräsentationen, *LILOG-Report 65*, IBM Deutschland (Oktober 1988).

[35] D. Reisberg and D. Chambers, Neither pictures nor propositions: The intentionality of mental imagery, *Proc. 8th Intern. Conf. of the Cognitive Science Soc.*, Amh. Mass. (1986) 208–222.

[36] Y. Shoham, What is the frame problem? in: P. Georgeff and A.L. Lansky, ed., *Proc. Reasoning about Actions and Plans* (Morgan Kaufmann, Oregon, 1986) 83–98.

[37] A. Sloman, Afterthoughts on analogical representations, *Proc. Theoretical Issues in Natural Language Processing*, Cambridge, MA (1975) 164–168.

[38] E.E. Smith and D.L. Medin, *Categories and Concepts* (Harvard University Press, Bostin, MA, 1981).

[39] E. Sober, Mental representations, *Synthese* 33 (1976) 101–148.

[40] G. Stanfill and D. Waltz, Toward memory-based reasoning, *Communication of the ACM* 29 (1986) 1213–1228.

[41] L. Steels, Steps towards common sense, *Proc. ECAI-88*, München (1988) 49–54.

[42] J.K. Tsotsos, A complexity level analysis of immediate vision, *International Journal on Computer Vision* 1 (1988) 303–320.

[43] J.K. Tostsos, J. Mylopoulos, H.D. Covvey and S.W. Zucker, A framework for visial motion understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2 (1980) 563–573.

[44] S. Ullman, Visual routines. *Cognition* 18 (1984) 96–159, also in: Steven Pinker, ed., *Visual Cognition* (MIT Press, Cambridge, MA, 1985).

[45] D.L. Waltz and L. Boggess, Visual analog representations for natural language understanding, *PRoc. Int. Joint Conf. on Art. Intell. IJCAI-6* (1979) 926–934.