Zusammenfassung

In diesem Bericht stellen wir die wesentlichen Bestandteile eines konzeptuellen Rahmens für Höhere Bilddeutung vor. Höhere Bilddeutung umfaßt definitionsgemäß alle Aufgaben der Szeneninterpretation "oberhalb" von Objekterkennung, also z.B. das Erkennen von Objektkonfigurationen, Vorgängen und absichtsvollen Handlungen. Die Modelle, auf die sich solche Interpretationen stützen, beschreiben im allgemeinen Aggregate aus bedeutungstragenden Teilen, die in zeitlichen sowie räumlichen Beziehungen zueinander stehen. Als Repräsentationsform werden Frames vorgeschlagen, ähnlich den Aggregatebeschreibungen aus der Konfigurierungstechnologie. Der Interpretationsprozess basiert auf dem Paradigma des Hypothetisieren-und-Testens und wird am Beispiel einer Tischdeckszene erläutert. Hypothesenbildung erfolgt im wesentlichen durch Teil-Ganzes-Schließen. Zur Evaluierung qualitativer zeitlicher Beschränkungen wird ein zeitliches Beschränkungsnetz vorgeschlagen. Ein ähnlicher Ansatz wird für räumliche Beschränkungen skizziert, die in Form von Gitterbereichen in objektbezogenen Referenzsystemen repräsentiert werden.

A Conceptual Framework for High-level Vision

Bernd Neumann
CSL, Hamburg University
neumann@informatik.uni-hamburg.de

Abstract

In this report we present essential elements of a conceptual framework for high-level vision (HLV). The scope of HLV is defined as scene interpretation above the level of object recognition. It is shown that models, on which such interpretations can be based, typically describe aggregates composed of meaningful parts, related to each other by temporal and spatial constraints. A frame-based representation is proposed which is based on techniques imported from configuration methodology. The hypothesise-and-test interpretation process is described for a table-laying example. It is shown that expectations are generated by part-whole reasoning. A temporal constraint net is proposed for the incremental evaluation of qualitative temporal constraints. A similar approach is sketched for spatial constraints which are represented by grid locations in a reference frame attached to an object.

Scope of high-level vision (HLV)

In this section we review developments in Computer Vision which contribute to a wider understanding of the vision task as compared to classical vision tasks such as recognizing or tracking single objects. These developments have to be taken into consideration when designing the conceptual framework for HLV in a cognitive agent as envisioned in the project CogVis.

From human vision it is evident that what we see is interpreted in the light of diverse knowledge and of experiences about the world. The scope of this knowledge - often termed common-sense knowledge - can best be seen when we consider silent-movie watching as a Computer Vision task, for example, watching and understanding a film with Buster Keaton. If a vision system were to interpret the visual information of such a film in a depth comparable to humans, the system would have to resort to knowledge about typical (and atypical) behaviour of people, intentions and desires, events which may happen, everyday physics, the necessities of daily life etc. This is knowledge far beyond the visual appearance of single objects, and a vision system capable of silent-movie understanding clearly has to solve tasks beyond single-object recognition.

As early as 1955 Computer Vision has been proposed as a task integrated in a cognitive context [Selfridge 55] and interacting with other cognitive processes. But Computer Vision research was in its infancy then, and a much narrower view of the vision task had to be pursued for several decades. The idea of integrating vision with other cognitive processes was actively investigated for the first time in the eighties in projects dealing with natural-language descriptions of imagery [vHahn et al. 80, Nagel 88, Neumann 89]. One of the important insights of this work was that qualitative descriptions had to be derived from geometric scene descriptions as an interface to language and symbolic reasoning.

An important paradigm which takes a more comprehensive view of the vision task is active vision [Bajcsy 88] (or purposive vision [Aloimonos 90]). In active vision the goal of vision is determined by the specific task which an agent may want to carry out. Active vision has been proposed as a departure from Marr´s view of vision as a general scene description task. Notions such as focus of attention, top-down control and vision-as-process are tied to the active-vision paradigm.

As Computer Vision was increasingly investigated in connection with actions in the real world, e.g. traffic behaviour, it became evident that spatial and temporal reasoning also played a part [Nagel 99, Neumann 99, Fernyhough et al. 98]. Spatial and temporal reasoning have been investigated in AI independently of vision for a long time, and reasoning services have been proposed, not all of which are useful for Computer Vision. More recently, description logics, extended by so-called concrete domains, have been shown to offer interesting support for vision tasks, especially for high-level tasks where complex spatial or temporal relations play a part [Moeller et al. 99].

While space and time are clearly essential domains for visual reasoning, knowledge representation and reasoning services must be considered as more common-sense knowledge is made available for vision tasks. Taxonomical relations between conceptual object descriptions, for example, can be exploited in recognition strategies, or automatic classification services which may be invoked as part of a knowledge representation system. Clearly, interfacing Computer Vision methodology with knowledge representation and reasoning is an important asset for comprehensive vision systems.

In the last decade, probabilistic models and learning have gained increased attention in the vision community. Visual behaviour can be predicted from a spatio-temporal context, models can be determined from the statistics of a large number of observations [Fernyhough et al. 98]. So far, learning is usually considered as a separate task, not integrated into vision systems. But as learning know-how from AI and Cognitive Sciences [Gärdenfors 00] is amalgamated into Computer Vision, it becomes conceivable to integrate model-building and experience-based vision into vision systems.

There are other AI topics besides knowledge representation and learning which have to be considered for HLV. As the term "purposive vision" suggests, planning and plan recognition is one such topic. In AI, a plan is a partially ordered set of actions designed to transform an initial world state into a goal state. From a vision point of view, knowledge about goals and plans to reach such goals support expectations about the development of a scene and hence provide useful top-down information for visual analysis.

As we investigate a conceptual framework for HLV in CogVis, we do not want to exclude any aspect which may contribute to vision. Hence we choose to define HLV as the part of Computer Vision which deals with scene interpretation "above" the level of object recognition. We will speak of a high-level interpretation of a scene, if it provides a meaningful description, typically in terms of qualitative abstractions, based on a larger spatial and temporal context.

As an illustrative example consider a street scene showing garbage collection. There is a garbage collection truck standing on the road and garbage bins standing on the curb. Men bring garbage bins to the rear of the truck, the bins are lifted, then lowered again and brought back to the curb. The men climb onto the truck and the truck moves on.

Let us assume that the following natural-language description of the scene is given : "There are men emptying garbage bins into a garbage-collection truck". This interpretation is typical for HLV and exemplifies several of the characteristics addressed above:

- The interpretation describes the scene in qualitative terms, omitting details.
- The interpretation may include inferred facts, unobservable in the scene.
- The scene is composed of several occurrences which contribute to the overall interpretation.
- Partial occurrences are spatially and temporally related.

The CogVis team at CSL in Hamburg prefers controllable indoor scenes for experimentation and has chosen a table-laying scene as a guiding example. Observed by stationary cameras, a human agent places covers onto a table. One task of the vision system is to recognise place-cover occurrences in an evolving scene based on a model containing generic knowledge about placing covers. A second task is to generate a place-cover model and other interesting models from repeated observation of reoccurring patterns in many scenes. This task addresses the issue of a vision memory and learning.

Basic framework for high-level scene interpretation

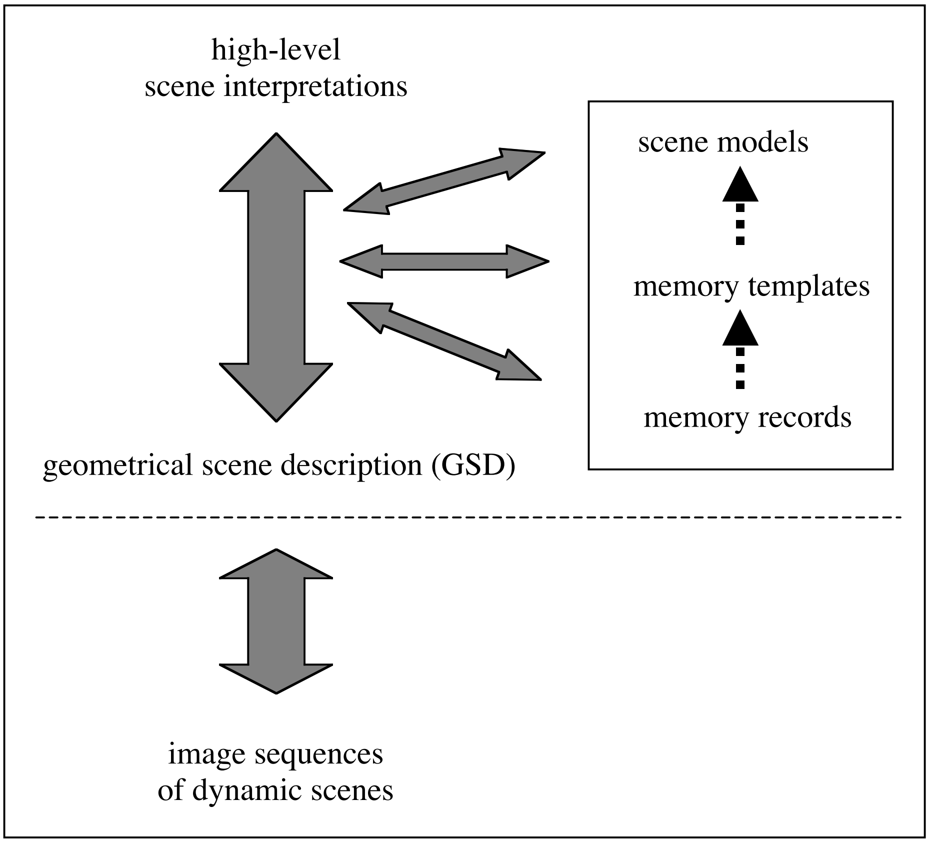

Figure 1: Basic building blocks for high-level scene interpretation

The main representational units and processes for high-level scene interpretation are shown in Figure 1. We will first describe the representational units.

A dynamic scene is a spatially and temporally coherent time-varying section of the real world.

With image sequences we mean image sequences taken from the scene using one or more cameras. In our work, there will be 3 stationary cameras observing a dynamic scene.

Geometric scene description (GSD) is the term for a quantitative object-level scene interpretation in terms of recognised objects and their (possibly time-varying) locations in the scene. A GSD is assumed to be available as input for HLV.

A scene interpretation is a scene description in terms of instantiated scene models, e.g. meaningful object configurations, occurrences, episodes and purposive actions. Scene interpretations are typically multi-level, with higher-level interpretations based on intermediate-level interpretations, and so on.

Both, GSD and interpretations, constitute the interpretation base. At a given time, the interpretation base contains the current description of the scene. It may also contain hypotheses which may not be upheld.

The block in the right half of Figure 1 is the model base. It contains recorded experiences and generic descriptions derived from experiences.

Memory records are stored copies of past scene interpretations. They are the data from which generic structures may be derived by learning processes.

Memory templates are generalised substructures of memory records, derived from reoccurring patterns in memory records. They may be the result of discovery and learning processes operating on a large number of memory records.

Scene models are conceptual entities for high-level scene interpretations. They may be constructed by knowledge engineers or result from supervised learning processes.

The basic processes are indicated by the arrows in Figure 1. Interpretation is performed by a hypothesise-and-test procedure. Hypothesis generation may be controlled by various factors: high-level expectations based on current hypotheses about ongoing occurrences, actions and goals, the current scene description in terms of the GSD, measures of saliency, likelihoods based on past experiences etc. Hypotheses are tentative instantiations of scene models or memory templates taken from the model base. Hypothesis test is a process aiming to verify that all elements of a scene model are consistent with the scene and enjoy sufficient evidential support.

The term expectation generation is used generally for a process which provides likelihoods for missing evidence based on established evidence. Expectation generation in HLV is essentially performed by hypothesis generation and part-whole-reasoning. Established evidence (elements of the GSD and established interpretations) are found to be part of a larger structure represented by a scene model. By hypothesising a scene model, one gives rise to expectations about other parts of the structure.

Temporal predictions play a special part if the scene is interpreted incrementally in real-time (or simulated real-time). This means that the scene is still unfolding while it is already being interpreted based on the available data. In this case, as time progresses, new evidence may confirm, contradict or extend predictions. Note that the process of checking current hypotheses in the light of new information is essentially the same hypothesis verification process which is part of the standard interpretation procedure.

Besides scene models, part-whole reasoning (and as such, expectation generation) may also exploit memory templates and memory records. Memory templates have a similar function as models, but they are nameless patterns, generated from experiences. Memory records, of course, are concrete past interpretations, parts of which may match the current situation and thus provide expectations.

The dotted arrows in the memory base in Figure 1 stand for learning processes. Learning is assumed to take place off-line, independently of the interpretation processes. The arrows indicate that memory templates and models may be generated from memory records. Learning processes will not be described in this report.

Choice of representational formalism

In this section we present the basic ingredients for representing scene models and interpretations in a declarative fashion. For HLV, the choice of a representational formalism is a choice of knowledge representation. Since knowledge representation is one of the oldest and most advanced fields of AI, there are numerous techniques available, and the question of how to represent knowledge for HLV is more a question of educated selection, guided by the requirements of HLV, than of creative design.

The following formalisms appear to be relevant:

- Frame representations are object-centered, expressive and most commonly used in AI.
- Relational structures provide graphical visualisations of relations and matching procedures.
- Description logics are attractive because of well-founded reasoning services.
- Constraints are useful for incremental evaluation of relations.
- Bayes nets come into play as probabilistic models.
- Neural networks provide associative structures and a cognitive learning paradigm.

We choose frame-based representations for HLV by reasons described below. But it is a useful research strategy to view representational issues from multiple perspectives. In particular, we will also take the description logic viewpoint and use the conceptual language of the description logic system RACER [Haarslev 01] to paraphrase the frame-based models. The idea is to invoke well-founded reasoning procedures for interpretation, prediction and learning [Neumann and Schröder 96]. This will be the subject of a forthcoming paper.

Above anything else, HLV models require the representation of aggregates, that is, of structures composed of parts with relations between the parts. Aggregates are common structures in other modelling systems, in particular in configuration systems [Cunis et al. 87]. Aggregates give rise to a partonomy which is the hierarchical structure induced by part-of relations.

Models will be more or less specific. For example, a model may describe an action where an agent (a hand) transports an arbitrary object from one place to another. Another model may describe a very specific transportation act where a saucer is transported onto a table. The latter may be represented as the specialisation of the former, or, expressed the other way around, the more general model subsumes the more special model. This relation between models generates another hierarchy, called subsumption hierarchy or taxonomy. Taxonomies are common in object-based representation systems.

Model definitions usually involve restrictive relations, e.g. spatial and temporal relations. In previous work on HLV interpretation [Neumann 89, Macworth 96, Neumann 97] and monitoring [Kockskämper 95], it has been shown that it is useful to model relations as constraints. Constraint systems support incremental evaluation and stepwise hypothesis instantiation. This is another aspect which HLV has in common with configuration methodology [Syska et al. 88] where constraints represent inter-object relationships.

We will describe models (henceforth also called concepts) in terms of frames which contain the following information:

- concept name
- taxonomical parent concepts
- parts
- constraints between parts

Parts play a special role as they constitute the components which make up an aggregate. They are also subject to the constraints expressed in the constraints section of the frame.

The following example frame describes an occurrence of the type "place-cover".

name:	place-cover
parents:	:is-a agent-activity
parts:	pc-pl :is-a plate
	pc-sc :is-a saucer
	pc-cp :is-a cup
	pc-tt :is-a table-top
	pc-tp1 :is-a transport with (tp-obj :is-a plate)
	pc-tp2:is-a transport with (tp-obj :is-a saucer)
	pc-tp3 :is-a transport with (tp-obj :is-a cup)
	pc-cv :is-a cover
time marks:	pc-tb, pc-te :is-a timepoint
constraints:	pc-tp1.tp-ob = pc-cv.cv-pl = pc-pl
	pc-tp2.tp-ob = pc-cv.cv-sc = pc-sc
	pc-tp3.tp-ob = pc-cv.cv-cp = pc-cp
	pc-cv.cv-tb ≥ pc-tp1.tp-te
	pc-cv.cv-tb ≥ pc-tp2.tp-te
	pc-cv.cv-tb ≥ pc-tp3.tp-te
	pc-tp3.tp-te ≥ pc-tp2.tp-te
	pc-tb ≤ pc-tp1.tb
	pc-tb ≤ pc-tp2.tb
	pc-tb ≤ pc-tp3.tb
	pc-te ≥ pc-cv.cv-tb
	pc-te ≥ pc-tb + 80Δt

Figure 2: Conceptual model for a place-cover occurrence

The place-cover model lists parts and constraints which must be fulfilled in a scene where a cover is being placed on a table.

The parts section lists local names and concept memberships of the essential visual phenomena constituting the place-cover occurrence. In particular, there must be 4 object instances of type plate, saucer, cup and table-top, 3 transport instances involving a plate, a saucer and a cup, and an instance of a cover configuration (concept definition not shown). Note that any entity associated with the occurrence place-cover may be listed as a part. Parts are assumed to provide partial evidence and will be the entry points for part-whole reasoning.

Following the parts section, the place-cover frame names the time marks which describe the beginning and ending of the occurrence. Time marks are subject to temporal constraints.

Finally, the constraints section establishes relations between the parts and between constituents of the parts. First, there are constraints establishing the identity between entities. Second, there are temporal constraints in terms of inequalities between time points which mark the beginning (tb) or ending (te) of an occurrence. Temporal constraints will be evaluated incrementally in a dedicated constraint net. The notion of an occurrence will be introduced in detail in the next section.

Other models, e.g. the model for a cover configuration, will also contain spatial constraints. A spatial constraint system is still in development.

Geometrical scene description and primitive occurrences

In this section we describe the interface between lower-level vision processes and HLV.

The input for HLV has already been introduced in Section 2 in terms of the geometrical scene description (GSD). A GSD is defined as a quantitative object-level scene interpretation in terms of recognised objects and their (possibly time-varying) locations in the scene. Ideally, this means that all objects relevant for a high-level scene interpretation have been recognised and tracked in a 3D scene coordinate system. In this case, HLV can in fact be restricted to interpretations in terms of aggregates and qualitative abstractions, as motivated above. As all vision researchers know, a perfect GSD cannot be expected in realistic applications. There will be unrecognised objects, interrupted tracks due to occlusion or segmentation faults, uncertain 3D information etc. It is therefore important that HLV processes

- do not rely on complete data, and
- provide top-down information in support of lower-level vision.

These requirements will be met by the hypothesise-and-test control regime of our conceptual framework. Using part-whole-reasoning, contextual evidence will be exploited to fill in missing evidence. We try to follow Kender´s candid definition that "vision is controlled hallucination".

The first task of HLV is to map the quantitative data of a GSD into qualitative entities which may play a part in higher-level models. This is done in three steps as shown in Figure 3.

In the first step, perceptual primitives are computed. This is a potentially rich set of measurements which can be immediately determined from a GSD. Here, we restrict our attention to object configurations and consider perceptual primitives which provide measurements between reference features of objects.

Reference features of objects are:

- locations (center of gravity, corners, point surface markings, etc.)
- lines (edges, axes of minimal inertia, line surface markings, etc.)
- orientations (inate, motion-based, viewer-based)

Perceptual primitives are:

- distance
- angle
- temporal derivatives thereof

In general, it may not be feasable to compute distances and angles beween all pairs of objects. Saliency measures and focus of attention come into play when a selection has to be made. We will bypass this problem in this report and assume that all perceptual primitives are available which play a part in higher-level interpretations.

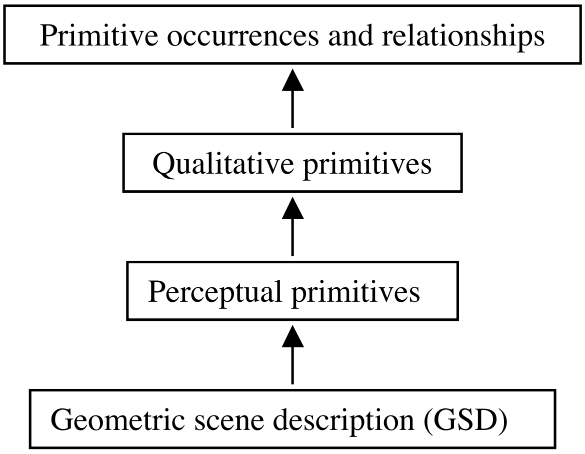
Primitive occurrences and relationships
Qualitative primitives
Perceptual primitives
Geometric scene description (GSD)

Figure 3: Providing primitives for high-level interpretations

In the second step, qualitative primitives are computed which are defined as predicates over perceptual primitives. The predicates can be characterised as discovering "qualitative constancies", e.g.

- constant values
- values within a certain range
- values smaller or larger than a threshold.

Spatial relations similar to natural language prepositions may be defined this way, e.g.

- degrees of nearness,
- directional sectors (e.g. front, rear, left, right),
- topological relations (e.g. within, overlap, outside),
- relative orientations (e.g. parallel, oblique, cross).

In dynamic scenes, temporal derivatives are particularly important, for example

- change of location (moving)
- change of orientation (turning)

- increasing, decreasing distance
- increasing, decreasing angle.

Qualitative primitives in dynamic scenes provide the basis for primitive occurrences. A primitive occurrence is defined as a conceptual entity where one or more objects give rise to a qualitative primitive which is true over a time interval. Typical primitive occurrences are:

- object motion,
- straight segment of an object motion,
- approach segment an object motion relativ to a second object,
- turning object motion,
- upward or downward motion.

The general representational form of a primitive occurrence is similar to scene models introduced earlier. For example, a straight-move occurrence is represented as shown in Figure 4.

name:	straight-move
parents:	:is-a move
parts:	sm-ob :is-a object
time marks:	sm-tb, sm-te :is-a timepoint
constraints:	sm-predicate

Figure 4: Conceptual model for a primitive straight-move occurrence

Note that for a concrete straight-move instance, sm-ob provides access to the quantitative location data of the GSD.

If a predicate over a perceptual primitive is true throughout a scene, one usually does not talk about an occurrence. We will use the term primitive relationship instead, well aware that there is no inherent representational difference between a constancy which happens to change within the duration of a scene and one which does not. Hence two stationary objects may be in an on-relationship (e.g. plate on table-top) or may be involved in an on-occurrence (e.g. plate on table-top from time 13 to 75).

Hypothesise-and-test cycle

In this section we describe hypothesis generation and testing, illustrated by an example. Figure 5 shows a snapshot of the emerging interpretation of a scene where an agent places a cover consisting of plate, saucer and cup, onto a table-top. In the right part of the figure, part of a model base is shown. The rest of the figure shows the interpretation base consisting of GSD data and hypotheses, connected by part-of links specified in the corresponding models.

For the sake of clarity, other links, for example :is-a links of the taxonomy and :instance links connecting instances and models, are not shown in the figure.

We will now sketch the interpretation process from the beginning of the scene up to the point shown in the figure.

We assume that the scene begins with a hand of an agent (agent1) moving together with a plate (plate1), while the other objects - saucer1, cup1 and table-top1 - are at rest. From the GSD, the primitive move occurrences move1 and move2 will be computed and updated as the scene evolves.

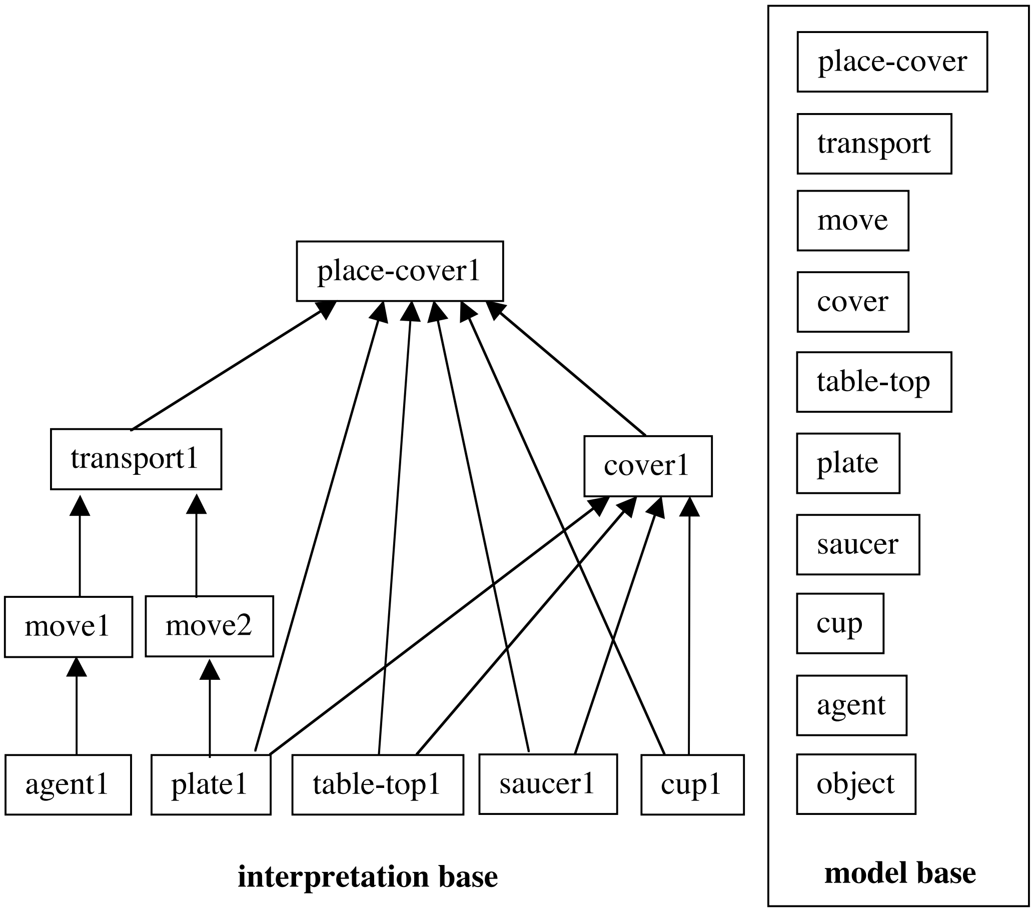

Figure 5: Emerging interpretation of a place-cover scene

Interpretation begins by selecting entries from the interpretation base which may be part of a higher-level structure. The selection is influenced by a focus control and by likelihood evaluations which will not be discussed in detail in this report. In our example, plate1, table-top1, saucer1 and cup1 are found to be possible parts of a cover (a cover model specifies the spatial relations between a plate, a table-top, a saucer and a cup). Hence a tentative cover1 hypothesis is created. Unfortunately, the spatial constraints specified by the cover model are violated as the objects are not yet in place. Hence the hypothesis will be removed (not shown in the figure).

A transport hypothesis is also found to be a likely candidate for a higher-level interpretation, since the interpretation base contains a moving hand and a moving object. The hypothesis is generated and can in fact be verified - for the time being - by evaluating the constraints of the transport model. For example, hand and object are verified to stay close to each other throughout the motion.

Note that the transport occurrence may not have terminated at this time. As long as motion1 and motion2 continue, the hypothesis may be falsified, if the continuing motion does not satisfy the constraints. On the other hand, evidence may be strong enough to permit useful predictions about the continued motion of hand and plate.

The interpretation process continues generating higher-level hypotheses based on the current interpretation. It is found that transporting a plate is part of the place-cover model, and a few other parts specified by this model are also present in the interpretation base (a saucer, a cup and a table-top). If this is considered strong evidence, a place-cover hypothesis will be generated.

The attempt to verify the place-cover hypothesis makes evident that important parts of the place-cover hypothesis - a saucer transport, a cup transport, and a cover - are still missing in the interpretation base. Temporal constraints may help to decide whether the missing parts may still be expected. In the place-cover model,we have a constraint placing the cup transport behind the saucer transport, constraints placing the beginning of a cover behind the end of all three transports, and a constraint limiting the overall duration of a place-cover occurrence. These constraints may still be satisfied at this time.

Expectations about missing parts may trigger corresponding hypotheses even if no evidence is present. Expectations may also influence lower-level processes, for example by providing a focus on areas where motion is expected.

As the scene evolves, additional hypotheses will be generated and incomplete hypotheses - such as the place-cover hypothesis - will be checked again. The temporal constraint net will monitor whether temporal constraints can still be satisfied and will signal a violation if, for example, the actual time exceeds the latest possible ending time according to the occurrence model.

The basic hypothesis-and-test procedure can be summarised as follows:

```
Hypothesise
        Select entries from the interpretation base.
        Determine model M which is supported by entries.
        Create hypothetical instance H of model M.
Verify
        Verify parts of H
                If part P of H is not in interpretation base then
                        Create hypothetical instance H´ of part P.
                        Verify H´
        Verify constraints of H
        If successful then add H to interpretation base.
```

This is, of course, only a rudimentary specification of the hypothesise-and-test cycle. It serves to illustrate the mixed bottom-up and top-down processing steps and the role of the constraints. It is evident that a sophisticated uncertainty management will be required to control this process. Uncertainty management will be based on the statistics which will be provided by the memory records of the vision memory.

Temporal constraint net

Temporal relations are modeled using a convex time point algebra [Vila 94]. The basic format of a (qualitative) temporal relation in this algebra is

t1 ≥ t2 + c12

where t1 and t2 are integer-valued time points and c12 is an integer-valued constant. Using such inequalities, it is possible to model important (but not all) features of the temporal structure of a scene model. In particular, one can express a convex subset of Allen´s interval relationships [Allen 83], for example

starts-within <=> int1.tb ≥ int2.tb + Δt, int2.te ≥ int1.te + Δt

with int1.te ≥ int1.tb + Δt, int21.te ≥ int2.tb

because of the interval property. Temporal relations of this kind can be concisely represented in a constraint net and efficiently evaluated. This has been shown in earlier work on event recognition [Neumann 89] and monitoring [Kockskämper et al. 94].

As an illustration, Figure 6 shows the constraint net for the temporal contraints of the place-cover model.

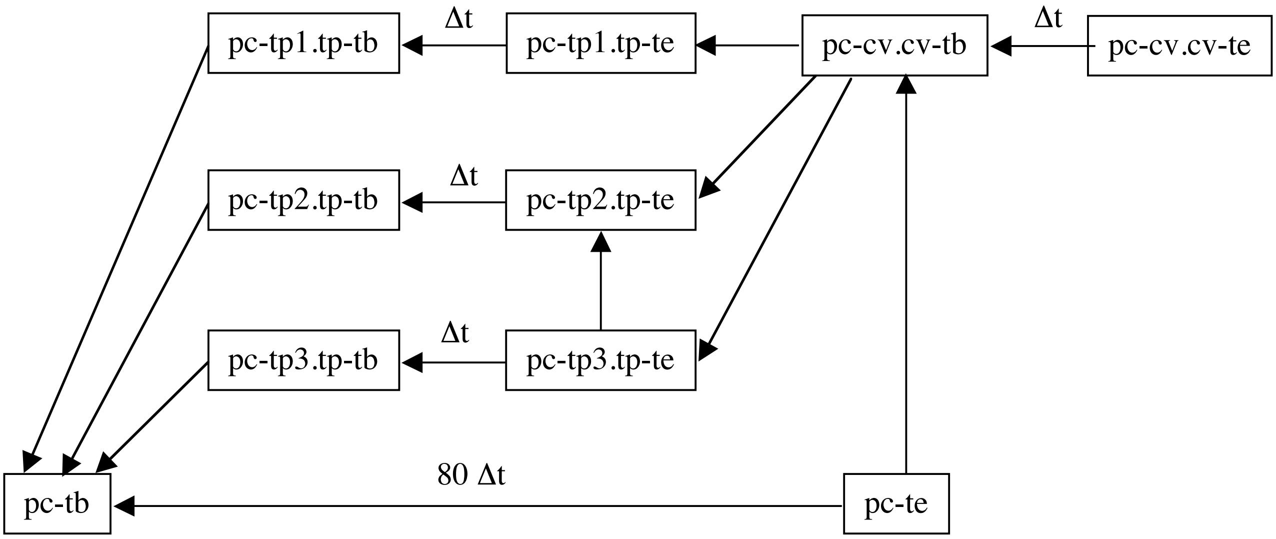

Figure 6: Temporal constraint net for the place-cover model

The nodes are timepoints corresponding to the time marks which make up the temporal structure of the model. The directed arcs represent inequalities, marked with an offset if the offset is different from zero. Each node is interval-valued, where the interval denotes the range of time points which is consistent with the constraints. Initially the intervals are open-ended, i.e. [-∞ +∞]. When an occurrence is selected as part of the model, the nodes corresponding to the time marks of that occurrence will receive concrete values. For example, if a transport-plate occurrence beginning at time 35 is selected, pc-tp1.tp-tb will receive the value [35 35]. New values are propagated through the constraint net, upper bounds along arrow directions and lower bounds against arrow directions, with offsets added or subtracted, respectively. The propagated values determine new interval boundaries if they constrain the old values. If the lower bound of an interval turns out to be larger than the upper bound, the constraints are inconsistent and cannot be satisfied with the selected instances.

It can be shown that each arc will only be traversed once, when a new value is propagated. Hence the complexity of this operation is $O(N^2)$ where N is the number of nodes.

Spatial relations and 3D structure

In this section we sketch our approach for representing spatial relations. This will also shed some light on the role of 3D information in our conceptual framework. Not all details are worked out yet as this is ongoing work.

Spatial relations must play a similar role in modelling and interpretation as temporal relations:

- They must be determined from quantitative data of the GSD.
- They must qualitatively constrain spatial relationships between parts of aggregates.
- They must allow incremental evaluation, as a model is instantiated step by step.

In addition, we want to determine typical spatial relations from visual experiences rather than using only predefined relations. Hence it must be possible to express qualitative spatial relations as a collection of concrete locations.

The basic idea is to represent a spatial relation as an assignment of possible locations in a reference grid attached to an object. We have chosen to consider mainly 2D spatial relations, but we allow the reference grid to be defined for any spatial orientation, depending on which spatial constraints have to be expressed. In our table-laying scenario, most spatial relations refer to a horizontal layout, hence the examples will involve reference grids parallel to the horizontal plane.

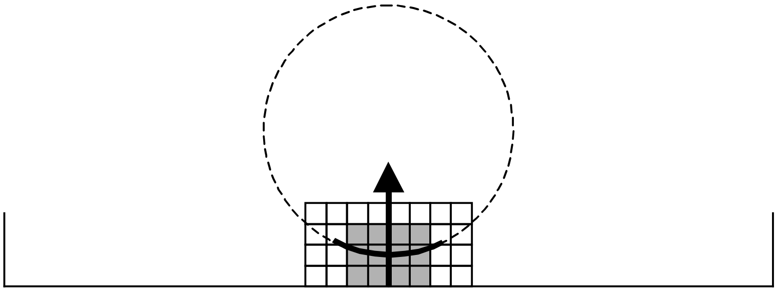

Figure 7: Reference grid for spatial relation between table border and plate

Figure 7 illustrates the principle. The constrained location of a plate border relative to a table border is defined as a set of grid cells (shaded) in a reference grid attached to the table border.

Figure 8 shows another example where grid cells are used to represent a sector oriented at an angle to the reference frame of the plate. One can see that also standardised spatial relations may be defined, for example corresponding to natural language terms such as in-front-of.

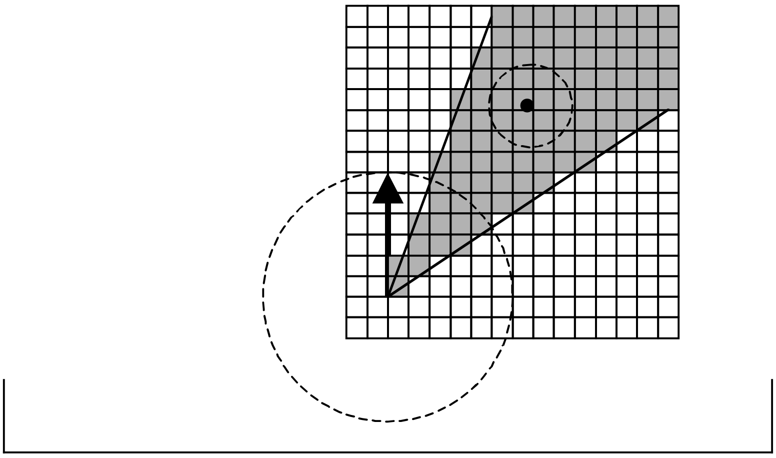

Figure 8: Reference grid for locations in a sector

As noted above, the reference grid of a model is oriented in a way which is most suitable to express the spatial relation in question. In order to verify whether one or more spatial relations of a model are fulfilled in a concrete scene, the reference grids of the model have to be mapped into a common reference frame which is typically an image plane to which GSD data refer. Here, spatial relations express constraints on possible locations of concrete objects of the current scene. Stepwise instantiation of a model will narrow down possible locations and provide guiding top-down information for lower-level processes.

Implementation and experiments

An implementation of a system for high-level interpretation of table-laying scenes is underway using the Common LISP Object System CLOS. Several modules described in this report are available from earlier work but need to be adapted. A simulator program has been implemented which generates experimental data of scenes with covers being placed on a table.

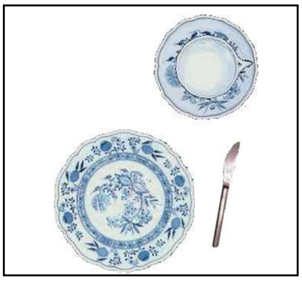

Figure 9: Cover configuration taken from a scene generated by a simulator

The simulator program will eventually be replaced by a multiple stationary camera setup ("smart room") and a low-level vision system providing the GSD. The work is part of the EU Project IST-2000-29375 CogVis.

References

[Allen 83]
Maintaining Knowledge About Temporal Intervals
J.F. Allen
in: Communications of the ACM 26 (11), 832-843, 1983

[Aloimonos 90]
Purposive and Qualitative Active Vision
J. Aloimonos
Proc. Image Understanding Workshop, 1990, 816-828

Badler 75
Temporal Scene Analysis: Conceptual Descriptions of Object Movements
N.I. Badler
Report TR 80, Dep. of Computer Science, University of Toronto, 1975

[Bajcsy 88]
Active Perception vs. Passive Perception
R. Bajcsy
Proc. of the IEEE, 76(8), 1988, 996-1005

[Cunis et al. 87]
Das PLAKON-Buch
R. Cunis, A. Guenter, H. Strecker (Eds.)
Informatik Fachberichte 266, Springer, 1987

[Fernyhough et al. 98]
Building Qualitative Event Models Automatically from Visual Input
J. Fernyhough, A.G. Cohn, D. Hogg
Proc. ICCV-98, IEEE Computer Society, 1998, 350-355

[Gärdenfors 00]
Conceptual Spaces - The Geometry of Thought
P. Gärdenfors
The MIT Press 2000

[Haarslev 01]
Description of the RACER System and its Applications
V. Haarslev
Proc. International Workshop on Description Logics (DL-2001), 2001

[Kockskämper et al. 94]
Extending Process Monitoring by Event Recognition
S. Kockskämper, B. Neumann, M. Schick
in: Proc. ISE-94, 455-460, 1994

[Kockskämper 95]
S. Kockskämper
Modeling and Prediction of Dynamic Behavior for Model-based Diagnosis
Proc. IEA/AIE-95, Melbourne, 1995, 285-304

[Mackworth 96]
Quick and Clean: Constraint-Based Vision for Situated Robots
A.K. Mackworth
in: Proc. ICIP-96, Vol. III, 1996, 189-792

[Moeller et al. 99]
Towards Computer Vision with Description Logics: Some Recent Progress.
R. Moeller, B. Neumann, M. Wessel
Proc. Speech and Image Understanding, IEEE Computer Society, 1999, 101-116

[Nagel 88]
From Image Sequences towards Conceptual Descriptions
H.-H. Nagel
Image and Vision Computing 6(2), 1988, 59-74

[Nagel 99]
From Video to Language - a Detour via Logic vs. Jumping to Conclusions
H.-H. Nagel
Proc. Speech and Image Understanding, IEEE Computer Society, 79-99

[Neumann 89]
Description of Time-Varying Scenes
B. Neumann
in: Semantic Structures, D. Waltz, Ed., Lawrence Erlbaum, 1989

[Neumann 97]
Providing Knowledge-Based Predictions for Dynamic Scene Analysis
B. Neumann
Proc. Workshop on Dynamic Scene Recognition from Sensor Data, Toulouse, Frankreich, 1997

[Neumann and Schröder 96]
How Useful is Formal Knowledge Representation for Image Interpretation?
B. Neumann, C. Schröder
in: Proc. Workshop on Conceptual Descriptions from Images, ECCV-96, 1996, 58 - 59

[Selfridge 55]
Pattern Recvognition and Modern Computers
O.G. Selfridge
Western Joint Computer Conf., 1955, 91-93

[Syska et al. 88]
Solving Construction Tasks with a Cooperating Constraint System
I. Syska, R. Cunis, A. Guenter, H. Peters, H. Bode
Proc. Expert Systems 88, Brighton, England, 1988

[Tsotsos et al. 80]
A Framework for Visual Motion Understanding
J.K. Tsotsos, J. Mylopoulos, H.D. Covvey, S.W. Zucker
IEEE PAMI-2, 1980, 563-573

[vHahn et al. 80]
The Anatomy of the Natural-Language Dialogue System HAM-RPM
W. v. Hahn, W. Hoeppner, A. Jameson, W. Wahlster
in: L. Bolc (Ed.): Natural Language Based Computer Systems, Muenchen, Hanser/McMillan 1980, 119-253

[Vila 94]
A Survey on Temporal Reasoning in Artificial Intelligence
L. Vila
AI Communications 7 (1), 4-28, 1994