Cognitive Vision - Towards Generic Models for Scene Interpretation

Bernd Neumann

Cognitive Systems Laboratory
Hamburg University
Germany

SFB/TR 8 Spatial Cognition Colloquium
Bremen University 2.12.2005

Agenda

- Challenges of Cognitive Vision
- Object Categorisation
- Scene Interpretation Framework
- Table-Setting Experiments

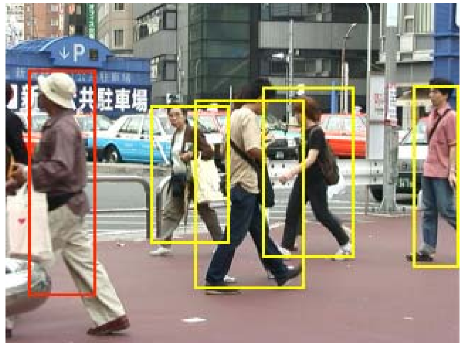
Cognitive Computer Vision

Cognitive computer vision is concerned with integration and control of vision systems using explicit but not necessarily symbolic models of context, situation and goal-directed behaviour. Cognitive vision implies functionalities for knowledge representation, learning, reasoning about events & structures, recognition and categorization, and goal specification, all of which are concerned with the semantics of the relationship between the visual agent and its environment.

Topics of cognitive vision:

- integration and control
- explicit models
- not necessarily symbolic
- context
- situation
- goal-directed behaviour
- knowledge representation
- learning
- reasoning
- recognition
- categorization
- goal specification
- visual agent

CogVis Topics (2001 - 2004)



- **Categorisation & Recognition of Structures, Events and Objects**

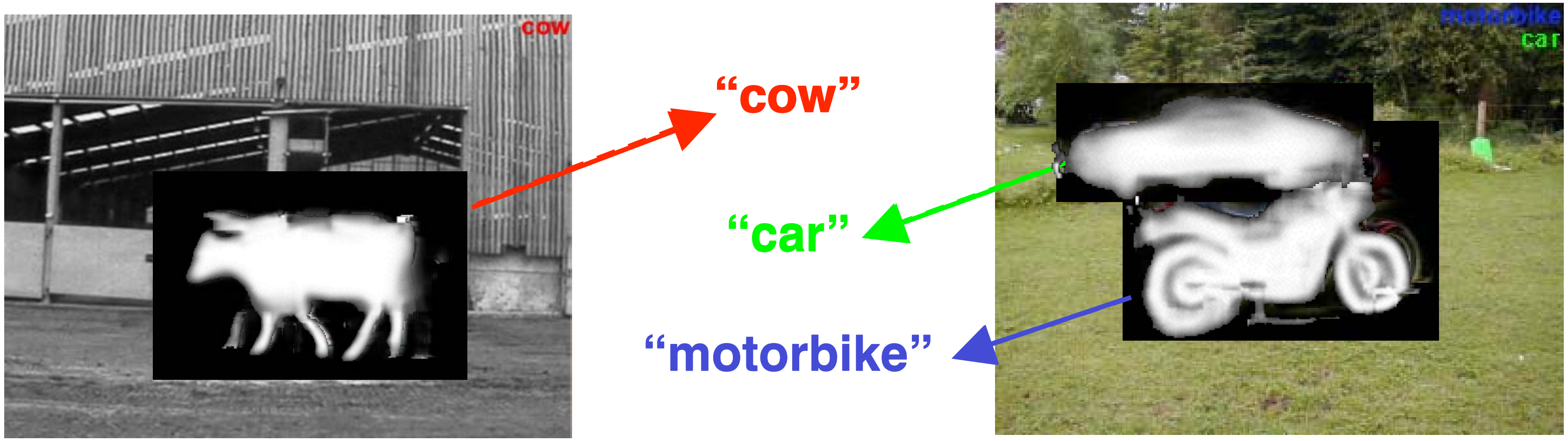
- **Interpretation and Reasoning**

- **Learning and Adaptation**

- **Control and Attention**

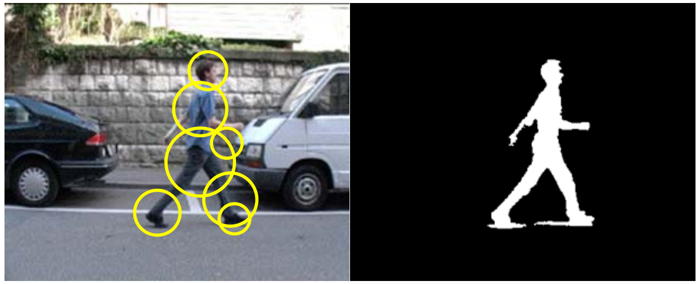
Object Categorization

Bastian Leibe, Bernt Schiele, Multimodal Interactive Systems, TU Darmstadt



- **Goals**
  - Learn to recognize object categories
  - Detect and localize them in real-world scenes
  - Segment objects from background

Implicit Shape Model - Representation

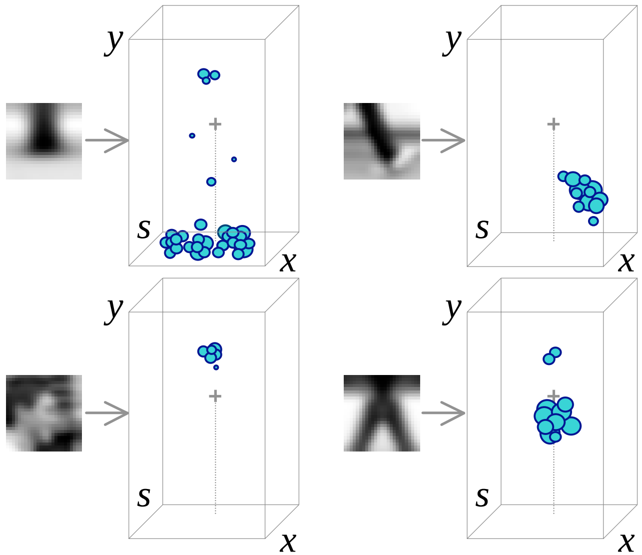


105 training images
(+motion segmentation)

Appearance codebook

- **Learn appearance codebook**
  - Extract patches at DoG interest points
  - Agglomerative clustering ⇒ codebook

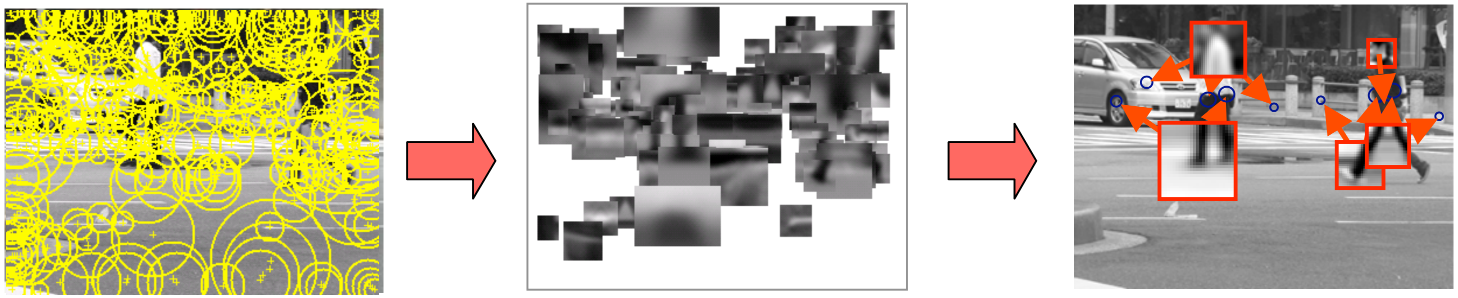
- **Learn spatial distributions**
  - Match codebook to training images
  - Record matching positions on object

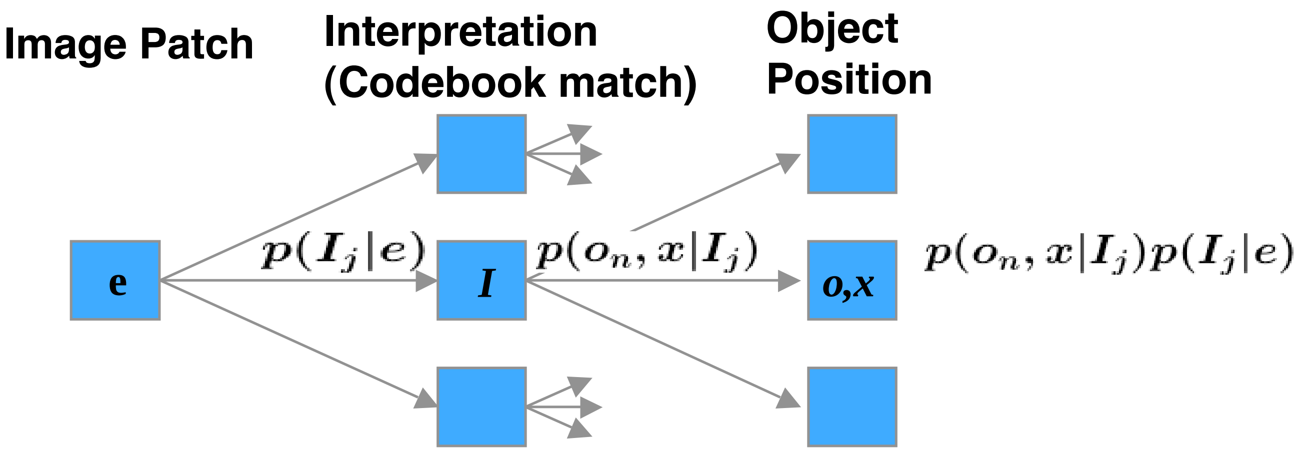


Spatial occurrence distributions

Implicit Shape Model - Recognition (1)

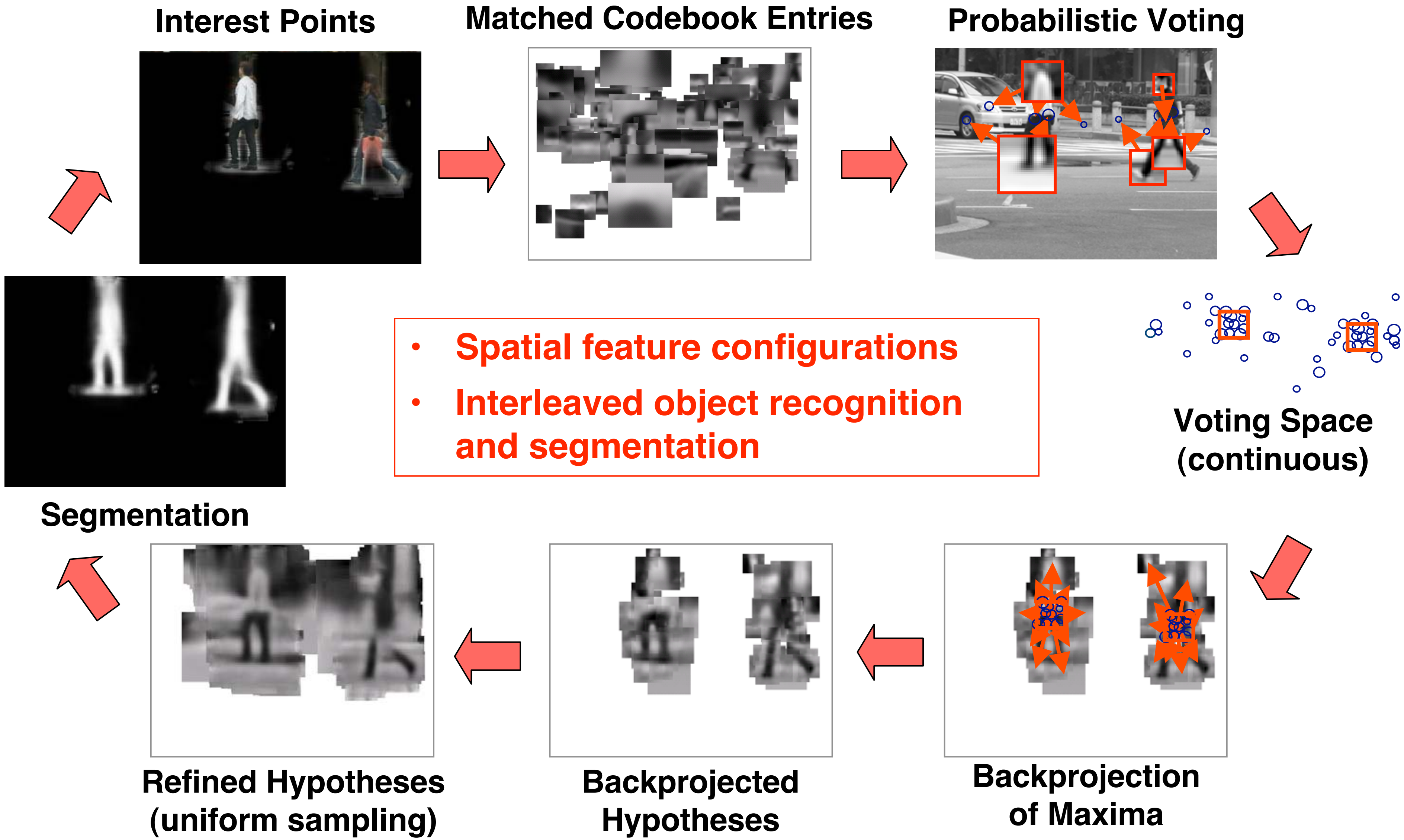
Interest Points | Matched Codebook Entries | Probabilistic Voting





$$p(o_n, x|e) = \sum_j p(o_n, x|I_j)p(I_j|e)$$

Implicit Shape Model - Recognition (2)



- Spatial feature configurations
- Interleaved object recognition and segmentation

Towards Scene Interpretation

garbage collection
+
mail delivery in Hamburg

unusual breakfast

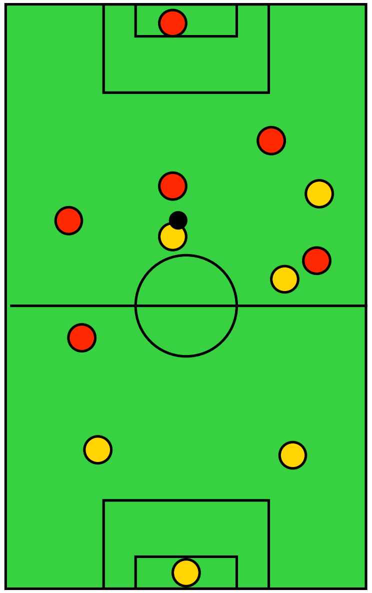
(Buster Keaton: The Navigator)

Challenges of Scene Interpretation

- Representing and recognizing structures consisting of several spatially and temporally related components (e.g. object configurations, situations, occurrences, episodes)
- Exploiting high-level knowledge and reasoning for scene prediction
- Understanding purposeful behaviour (e.g. obstacle avoidance, grasping and moving objects, behaviour in street traffic)
- Mapping between quantitative and qualitative descriptions
- Natural-language communication about scenes
- Learning high-level concepts from experience

Some Application Scenarios for High-level Scene Interpretation

- street traffic observations (long history)
- cameras monitoring parking lots, railway platforms, supermarkets, nuclear power plants, ...
- video archiving and retrieval
- soccer commentator
- smart room cameras
- autonomous robot applications
  (e.g. robot watchmen, playmate for children)



Towards Generic Models for Scene Interpretation

- Need for model-based approach
  - spatially and temporally coherent configurations
  - organising relevant knowledge
- Logic-based vs. probabilistic models
  - deduction, rules, uncertainty, consistency
- Interface to low-level vision
  - signal-symbol interface
  - quantitative-qualitative mapping
- Interpretation strategies
  - bottom-up vs. top-down
  - varying context
  - prediction

Conceptual Models for Scene Interpretation: Aggregates

aggregate name
parent concepts
external properties
parts
constraints between parts

➡ compositional hierarchies

➡ taxonomical hierarchies

representation by DL in principle possible

Occurrence Model for Placing a Cover

Composite occurrences are expressed in terms of simpler models

name:	place-cover
parents:	:is-a agent-activity
properties:	pc-tb, pc-te :is-a timepoint
parts:	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
constraints:	pc-tp1.tp-ob = pc-cv.cv-pl
	pc-tp2.tp-ob = pc-cv.cv-sc
	pc-tp3.tp-ob = pc-cv.cv-cp
	...
	pc-tp3.tp-te ≥ pc-tp2.tp-te
	pc-tb ≤ pc-tp3.tb
	pc-te ≥ pc-cv.cv-tb

Scene Interpretation as Model Construction

Construct a mapping of
- **constant symbols into scene elements D**
- **predicate symbols into predicate functions over D**

such that all predicates are true.

Operational semantics of low-level vision provide mapping into primitive constant and predicate symbols.

Finite model construction (Reiter & Mackworth, 87):

Domain closure and unique name assumption => problem can be expressed in Propositional Calculus and solved as a constraint satisfaction problem (CSP)

Partial model construction (Schröder 99):

- model may be incomplete, but must be extendable to a complete model
- disjunctions must be resolved

Practical Requirements for Partial Logical Models

- **Task-dependent scope and abstraction level**
  - no need for checking all predicates
  e.g. propositions outside a space and time frame may be uninteresting
  - no need for maximal specialization
  e.g. geometrical shape of "thing" suffices for obstacle avoidance

- **Partial model may not have consistent completion**
  - uncertain propositions due to inherent ambiguity
  - predictions may be falsified

- **Real-world agents need single "best" scene interpretation**
  - uncertainty rating for propositions
  - preference measure for scene interpretations

⇒ **Logical model property provides only loose frame for possible scene interpretations**

Stepwise Construction of Partial Models

Four kinds of interpretation steps for constructing interpretations consistent with evidence:

Aggregate instantiation
Inferring an aggregate from (not necessarily all) parts

Instance specialization
Refinements along specialization hierarchy or in terms of aggregate parts

Instance expansion
Instantiating parts of an instantiated aggregate

Instance merging
Merging identical instances constructed by different interpretation steps

Repertoire of interpretation steps allow flexible interpretation strategies
e.g. mixed bottom-up and top-down, context-dependent, task-oriented

Preferred Interpretation Steps

- **Logical framework may provide infinitely many partial models**
  e.g. involving objects outside the field of view
- **Wrong choices among alternative interpretation steps may cause severe backtracking**
  e.g. wrong part-whole reasoning

Probabilistic approach based on scene statistics:

Select interpretation steps which construct the most likely interpretation given evidence

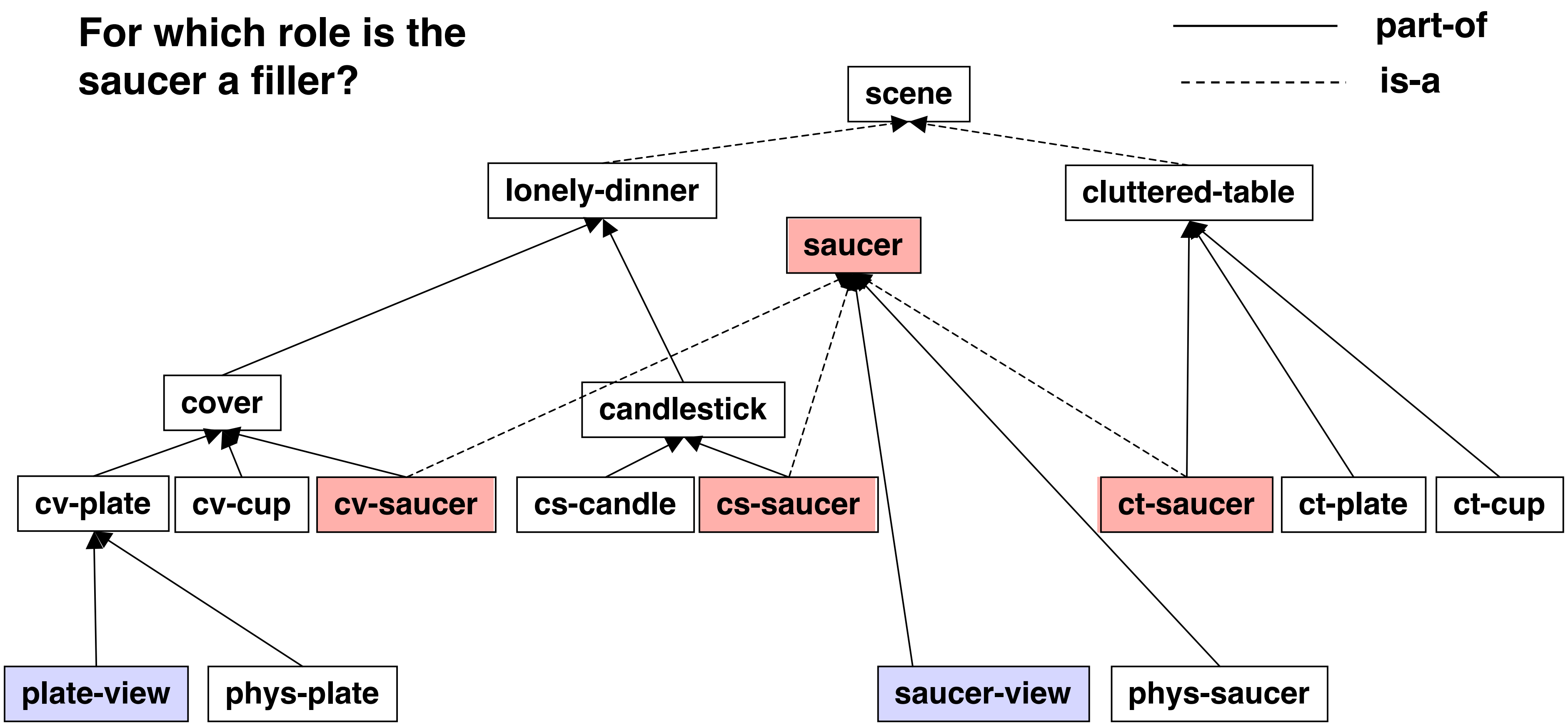
Probability distributions for

- **concept specializations**
  e.g. dinner-for-one vs. dinner-for-two
- **choices among individuals**
  e.g. choices of colours
- **discrete domain quantities**
  e.g. locations and time points

→ **multivariate distributions instead of constraints**

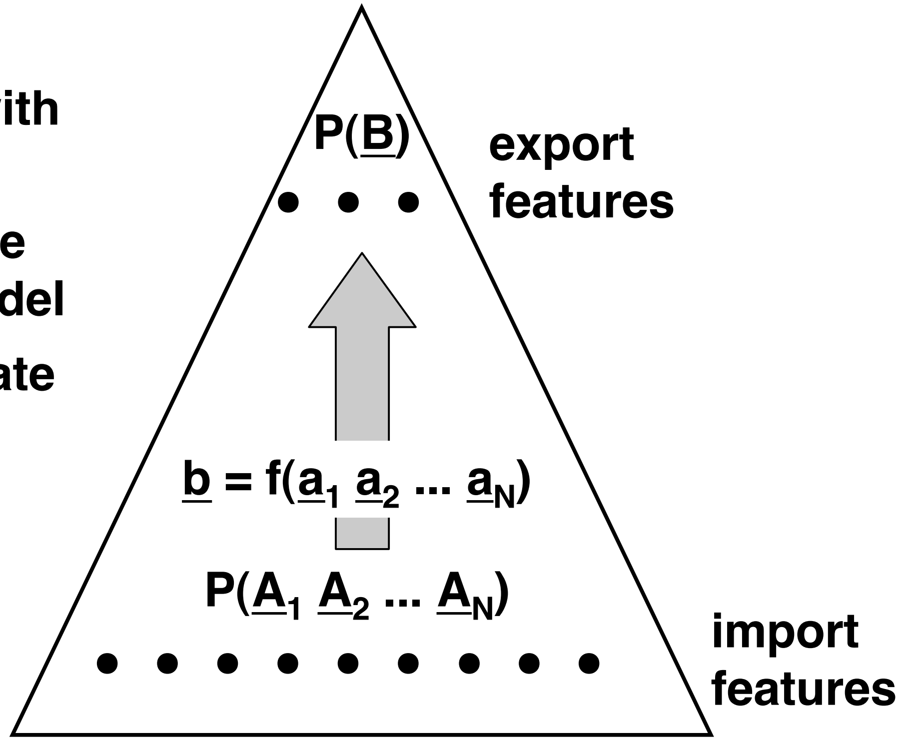
Example for Probabilistic Interpretation Decisions

For which role is the saucer a filler?



Integrating Bayesian Networks with DL Aggregates

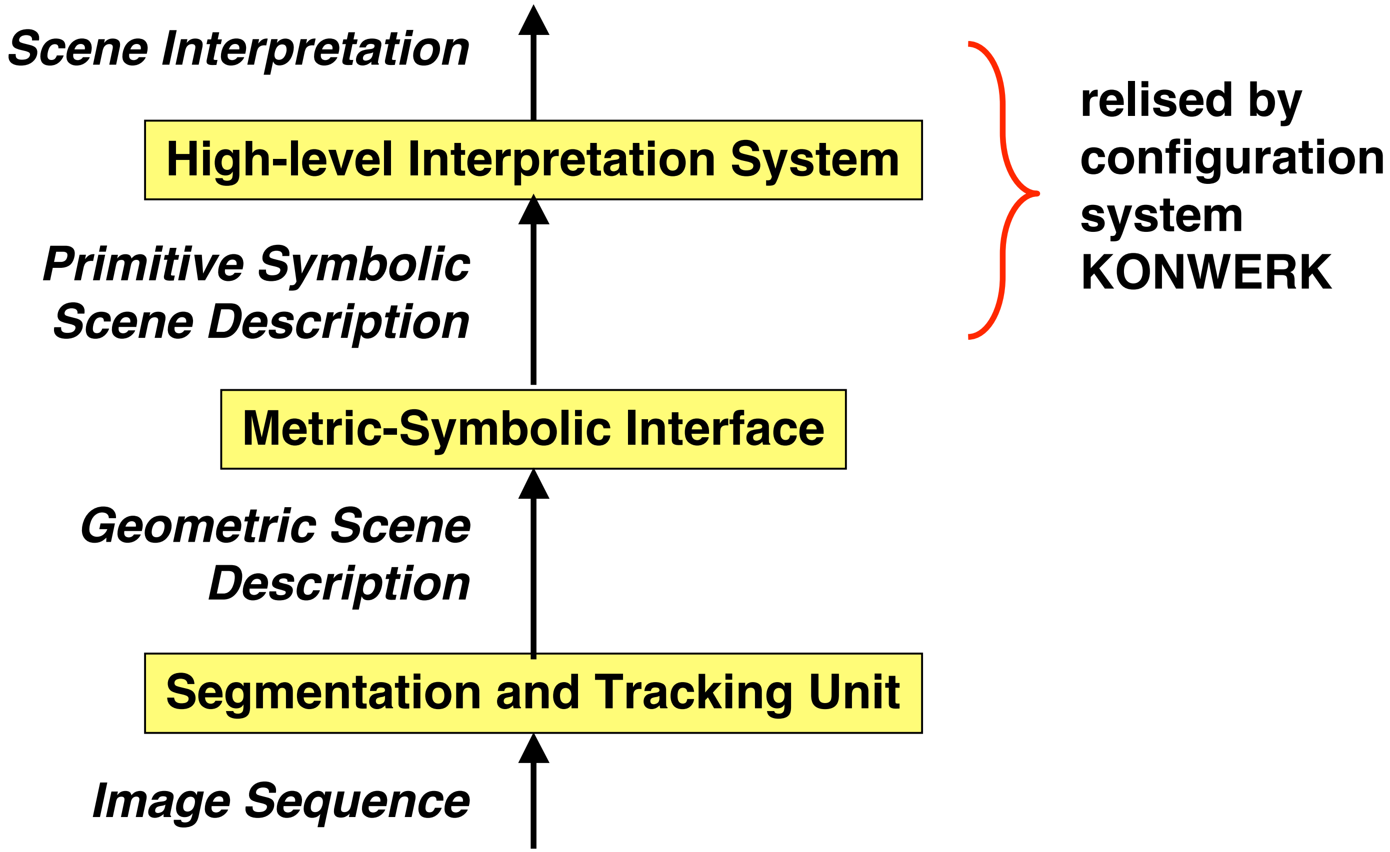
- Each aggregate is associated with a Bayes Net fragment
- An operational Bayes Net can be constructed for each partial model
- Abstraction property of aggregate fragments ensures efficient probability computations



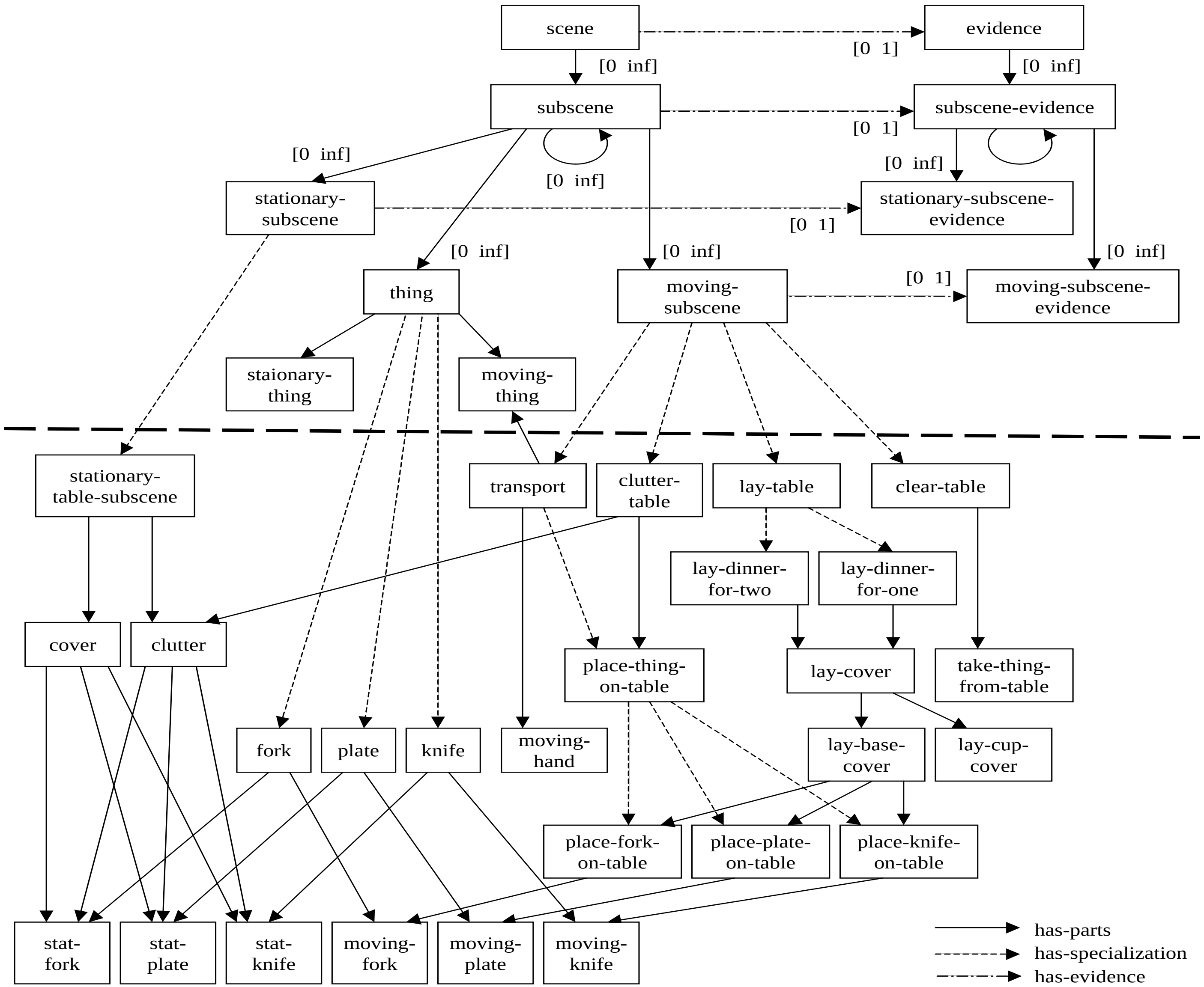
<u>Example</u>: Aggregate "cover"

JPD $P(\underline{A}_1\ \underline{A}_2 \ldots \underline{A}_N)$ for cover parts locations is mapped into JPD $P(\underline{B})$ for cover bounding-box location

Structure of Scene-Interpretation System SCENIC



Structure of Conceptual Knowledge Base of SCENIC

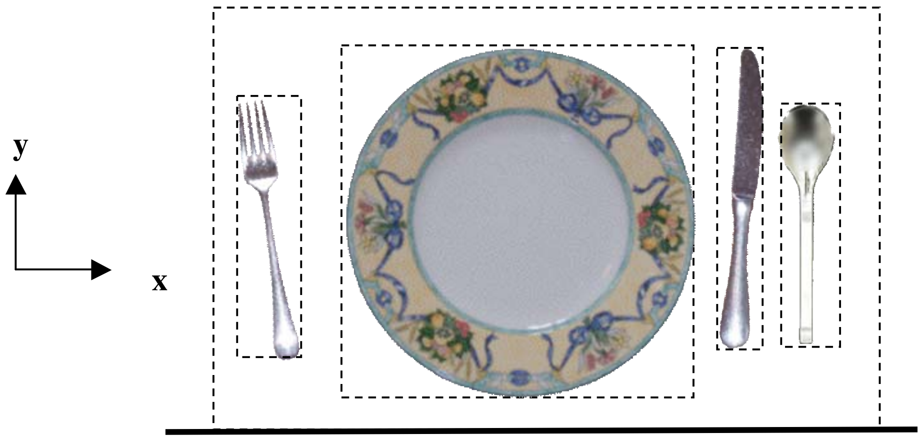


Example of Table-laying Scene

Stationary cameras observe living room scene and recognize meaningful occurrences, e.g. placing a cover onto the table.

In the following experiment: laying a table for a dinner-for-2

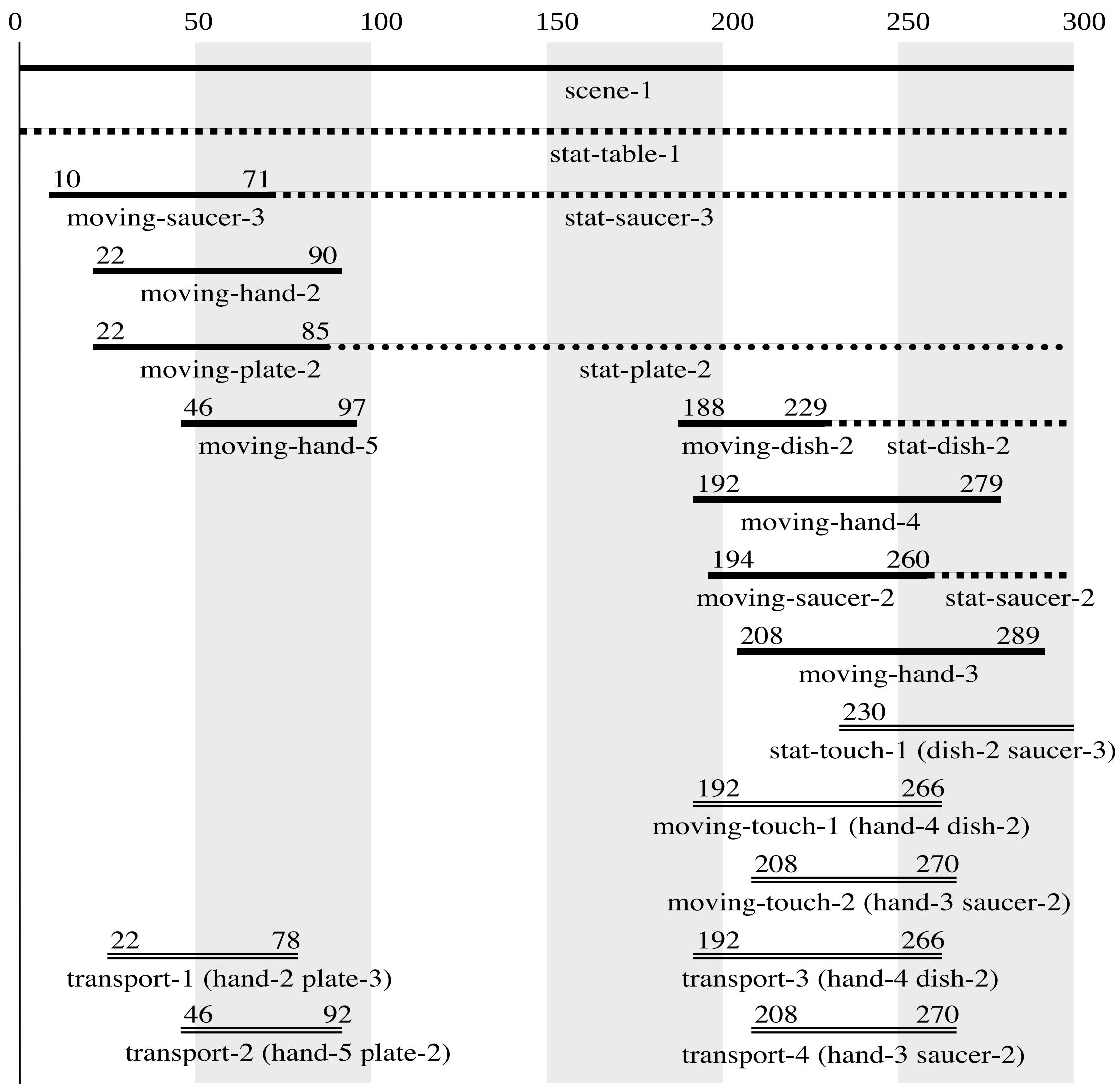
Bounding-Box Abstractions



- **object shapes are represented as 2D boxes**
- **aggregates hide internal structure**

- **box locations and distances are interval-valued**
- **value ranges and their correlations may be described by joint probability distributions**

Initial Bottom-up Instantiation of Concepts

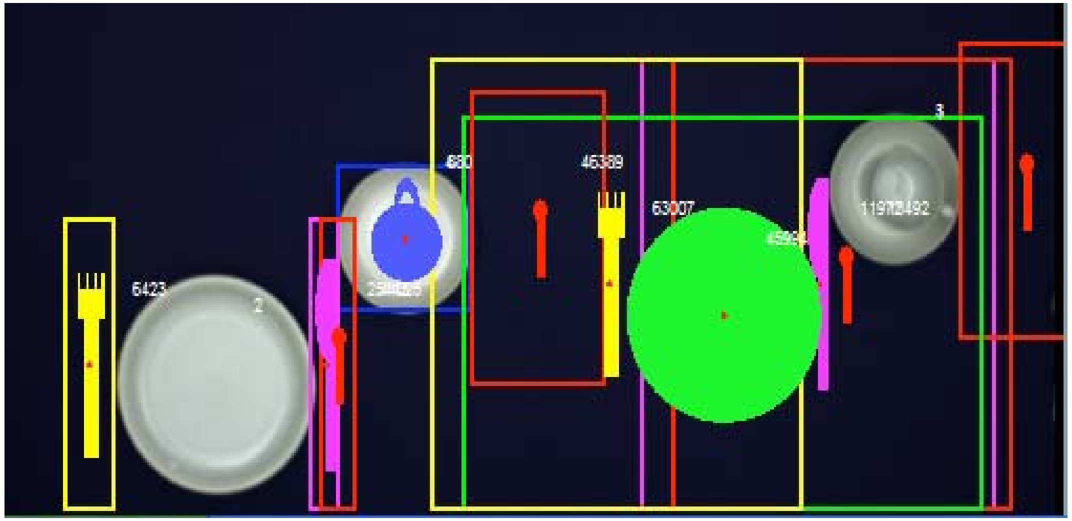


primitive stationary concepts

primitive motion concepts

aggregate concepts

Experimental Results (1)



natural views = evidence

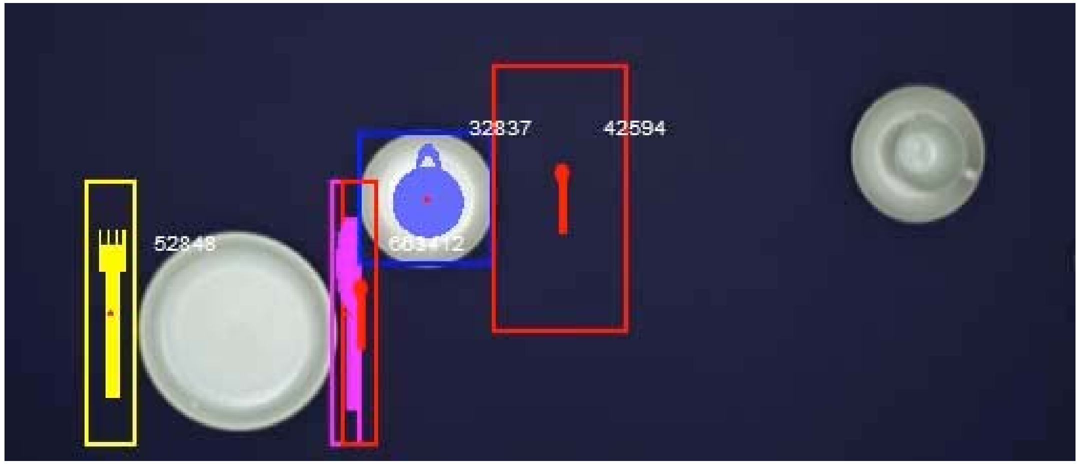
coloured shapes = hypotheses

boxes = expected locations

Intermediate state of interpretation after 51 interpretation steps:

- **"lay-dinner-for-2" hypothesis based on partial evidence**
- **predictions about future actions and locations**
- **high-level disambiguation of low-level classification**
- **influence of context**

Experimental Results (2)



- **alternative interpretation in terms of top-down choices "dinner-for-one" and "cluttered-table" (after backtracking)**

Conclusions

- Aggregates embedded in a compositional hierarchy are the key concepts for high-level scene interpretation
  - generic structure is based on components and relations between components
  - learnability is useful guide for conceptualisations

- Scene interpretation can be modelled equivalently as
  - partial logical model construction guided by a probabilistic preference measure
  - probabilistic inference constrained by logical consistency requirements