

# Towards an Architecture for Cognitive Vision using Qualitative Spatio-Temporal Representations and Abduction

A G Cohn, D R Magee, A Galata, D C Hogg, S M Hazarika

School of Computing, University of Leeds, Leeds, LS2 9JT, UK  
{agc,drm,afro,dch,smh}@comp.leeds.ac.uk.

**Abstract.** In recent years there has been increasing interest in constructing cognitive vision systems capable of interpreting the high level semantics of dynamic scenes. Purely quantitative approaches to the task of constructing such systems have met with some success. However, qualitative analysis of dynamic scenes has the advantage of allowing easier generalisation of classes of different behaviours and guarding against the propagation of errors caused by uncertainty and noise in the quantitative data. Our aim is to integrate quantitative and qualitative modes of representation and reasoning for the analysis of dynamic scenes. In particular, in this paper we outline an approach for constructing cognitive vision systems using qualitative spatial-temporal representations including prototypical spatial relations and spatio-temporal event descriptors automatically inferred from input data. The overall architecture relies on abduction: the system searches for explanations, phrased in terms of the learned spatio-temporal event descriptors, to account for the video data.

## 1 Introduction

There has been extensive research into techniques for Computer Vision (CV), but much of this has concentrated on important, but low level methods. Although these low level techniques can sometimes be applied directly in a system, in general, a more high level understanding of the scene will be required. The relative paucity of research in this area<sup>1</sup> has resulted in a number of EU funded projects on *Cognitive Vision* which allow a much greater “semantic” access to and processing of visual information. The University of Leeds is a partner in one such project, CogVis (Cognitive Vision Systems, IST-2000-29375). This paper describes our approach to the goal of creating a cognitive vision system, and in particular the combination of qualitative spatial reasoning techniques with more conventional CV research.

---

<sup>1</sup> Though see, e.g. [1–4].

First, it is worthwhile quoting from the Technical Annexe of CogVis to give a “definition” of cognitive vision:

“Considering the general definition of cognition as “the process of knowing, understanding and learning things” it is possible to derive some key characteristics for cognitive vision:

- Vision is a process that operates in a spatio-temporal context. I.e. vision is not instantaneous, it evolves over time and incorporates information to generate “answers”.
- Vision uses and generates knowledge (that includes information that is not organized spatially). This implies that a fundamental part of studies of visual processes is consideration of representations and memory.
- The visual process generates/maintains models of the environment in terms of its geometry, and semantic labels for events and entities in the environment. I.e. “understanding” implies an ability to generate an explicit description of the perceived world in terms of objects, structures, events, their relations and dynamics that can be used for action generation or communication.
- Learning implies an ability to generate open-ended models and representations of the world. That is, the model of the system and its use cannot be based on a closed world assumption, but rather on a model that allows automatic generation of new representations and models.
- Vision is a process which implies that it operates in the context of an “agent” that provides a task context and has finite resources in terms of computation, memory and bandwidth.”

Thus key to our approach will be the integration of learned models, of explicit spatio-temporal representations and open ended reasoning allowing the generation and explicit manipulation of symbolic hypotheses.

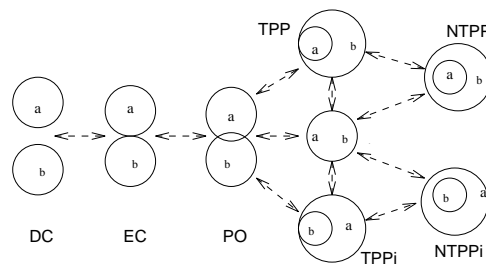
## 2 The Traditional Computer Vision Approach

Traditionally, CV approaches the problem of scene understanding as one of finding methods to transform between input images or sequences and an  $N$  dimensional parameterisation (where  $N$  is arbitrary but fixed) [5–9]. Ad-hoc methods are often used for representing and understanding multiple objects. Many spatial and temporal classification and prediction methods (e.g. Hidden Markov Models) have been developed based on this ‘world in  $N$  dimensions’ paradigm [10, 11]. The problem is that the world is not generally well described by an  $N$  dimensional parameterisation (although the higher the  $N$ , the better the description generally), rather as the sum of a number of concepts. Methods attempt to reduce the computational dimensionality without reducing the parameterisation dimensionality (for reasons of computational efficiency) by using dimensionality

compression techniques such as Principal Components Analysis or Vector Quantisation. This does not however solve the problem that a fixed  $N$  dimensional representation is not a good way of representing a general scene. CV methods generally *approximate* the real world and thus are rarely 100 % *accurate*.

### 3 Qualitative Spatio-temporal Representations and Reasoning

The development of Qualitative Spatial Reasoning (QSR) [12] has been driven by the realisation that much cognitive representation and processing of spatial data is qualitative – e.g. most everyday natural language spatial expressions are purely qualitative (“on the table”, “behind the tree”, “in the bottle”) and moreover that much uncertainty in spatial data can be abstracted away through the use of qualitative representations that discretize a continuous space into a finite and small number of *relevant* possibilities – qualitative representations are typically abstract but accurate (rather than precise and possibly inaccurate). Thus, for example, the RCC-8 calculus [13] has eight jointly exhaustive and pairwise disjoint relations categorising possible topological relations between a pair of regions – see figure 1; a very similar calculus has also been derived from alternative semantic primitives [14]. Indeed, the use of regions as a primitive spatial entity (rather than points) also helps abstract away from uncertainty. If the boundary of real world regions is unknown or in some other way indeterminate, then an extension of the calculus has been designed to handle such regions [15] (see also [16]). Other QSR calculi have been designed to represent and reason about orientation (e.g. [17–19]), convexity [13], shape (e.g. [20]) and congruence [21, 22]. An important notion when considering dynamic spatial knowledge is that of a *continuity network* or *conceptual neighbourhood* which specifies which relations are neighbours as objects move or transform continuously over time as this allows for prediction and explanation of spatio-temporal data – see figure 1. For a survey of QSR see [12] or [23].



**Fig. 1.** 2D illustrations of the relations of RCC-8 calculus and their continuous transitions (*conceptual neighbourhood*).

When reasoning with qualitative spatial data over time, one possibility is to take a ‘snapshot’ viewpoint, and describe dynamic behaviour as a set of temporal states, where each state consists of a qualitative spatial representation and their temporal relationship described by a temporal logic. This approach has been extensively investigated by [24–26] and a number of useful complexity results are given. An alternative approach is to view the world as spatio-temporal histories [27] and extend purely spatial qualitative representation languages to qualitative spatio-temporal languages with relations which hold between such space-time histories [28–30].

To apply QSR methods to CV requires a qualitative description of the world/scene as an input. In constructing this description, the system abstracts away from the initial (potentially erroneous) data and thus may remove error components in the data (e.g. by abstracting point locations to being within a certain region, and by choosing one of a small finite set of relations rather than exact, but inaccurate relation over real valued data). By comparison with a conventional CV approach, where approximation may lead to inaccuracies, a qualitative approach will typically be indefinite but accurate. This approach can deal with some modes of input error (e.g. additive noise) but may fail to deal with other error modes (e.g. missing data, erroneous extra data). The use of conceptual neighbourhoods in [4] can help deal with certain kinds of missing data by allowing interpolation of “missing” intermediate relations in an event sequence, or by using them to filter out noise. Conceptual neighbourhoods may also be used to help predict the next qualitative state (cf [31]).

## 4 A Logical Approach to Cognitive Vision

Computer Vision falls into a class of problems where some sensor data,  $\Omega$ , is acquired, and has to be interpreted relative to some already existing body of knowledge. Typically this body of knowledge falls into two categories: a very general, usually relatively domain independent knowledge base,  $\Sigma$ , and a more specific one,  $\Phi$ , which may depend much more on the task(s) at hand. The problem is to explain the sensor data  $\Omega$  given the prior knowledge. From a logical point of view, we can express this thus – what explanation  $\Delta$  makes the following statement true:

$$\begin{array}{c} \downarrow \downarrow \downarrow \\ \Sigma, \Phi, \Delta \models \Omega \\ \downarrow \end{array} \quad (1)$$

The arrows above the formula indicate that these items are ‘inputs’, whilst the arrow below the formula indicates that  $\Delta$  is the output – the abduced explanation.

This form of inference is called *abduction*. Shanahan [32, 33] has applied this form of inference to the problem of abducing maps from robotic (non video) sensor data – see also [34]. More recently, he has also applied this approach to robotic vision [35] where he proposes to use abduction to formally explain all

visual data either as a picture object or as noise, and preferring explanations with a higher “explanatory value”, i.e. which explain as little as possible as noise. The abduced explanations can then be used to feedback into the sensory action planning of the robot: it may initiate sensory actions to verify abduced hypotheses (e.g. by adjusting its noise thresholds, or even by attempting to touch or nudge a hypothesised object).

In the cognitive vision setting, we could view  $\Sigma$  as a background spatio-temporal theory (which might include, e.g. RCC-8),  $\Phi$  as a set of possible behaviour patterns expressed in  $\Sigma$  (e.g. being stationary, bouncing up and down, descending,...),  $\Omega$  is a qualitative abstraction of the spatio-temporal video data, and  $\Delta$  is a set of behavioural instances which make the entailment true (i.e. explain the observations).

This is the core of our framework: the entire cognitive vision system is driven by the need to abduce explanations of sensor data given prior background knowledge.

A major issue is where does the background knowledge  $\Sigma$  and  $\Phi$  come from? Very frequently, even in cognitive vision systems such as [3], this knowledge is explicitly programmed in by the human system builder. However this can be a tedious and often error prone process. Moreover, in a dynamic situation, or where the system is to be applied in different settings or contexts, different  $\Sigma$  and in particular  $\Phi$  will be required, which adds greatly to the difficulty of acquiring such knowledge.

In fact, just about any computer vision system can probably be viewed in this way, whether it is explicitly logic based or not. What distinguishes our proposed approach is that:

- we will explicitly build our system around the entailment (1), and will use logical inference methods.
- the candidate hypotheses are expressed in a (largely) qualitative spatio-temporal representation language.
- we will, as far as possible, acquire  $\Phi$  and possibly  $\Sigma$  too, automatically, as an *inductive* learning process from actual sensor data.

We argue that this has several advantages:

- The processes involved in interpreting visual data can be easily explained in terms of standard logical reasoning steps.
- Using a *qualitative* spatio-temporal representation allows easier generalisation of classes of different behaviours, guards against propagation of errors, and may facilitate the cognitive comprehension of spatio-temporal knowledge.
- Different variants of the architecture can be viewed as variants in control of the inferential process, different representation languages, different background knowledge bases, whilst maintaining a common architecture.
- The ability to learn the background knowledge makes the system much more robust and easier to field in different situations and domains.

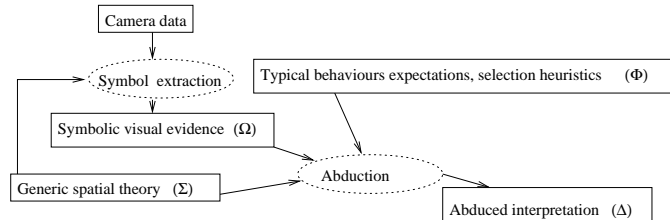
- The use of logic allows the explicit manipulation of alternative hypotheses and the general statement of background knowledge. It also allows partial and indefinite knowledge to be represented. Equally importantly, such knowledge might be reused in other tasks associated with an artificial agent such as planning or map building.

## 5 Our Existing Implementations

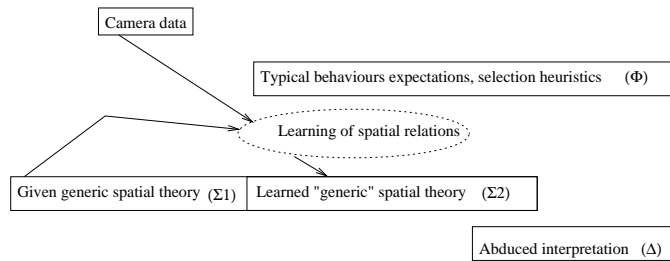
In [4], we have shown how qualitative spatio-temporal models of events in traffic scenes (e.g. following, overtaking) can be learnt. Using an existing tracking program which generates labelled contours for objects in every frame, the view from a fixed camera is partitioned into semantically relevant regions based on the paths followed by moving objects. The paths are indexed with temporal information so objects moving along the same path at different speeds can be distinguished. Using a notion of proximity based on the speed of the moving objects and a description of the relationship between close objects using QSR calculi for relative direction and relative orientation with respect to the path being travelled, event models describing the behaviour of pairs of moving objects can be built, again using statistical methods. The system has been tested on a traffic domain and learns various event models expressed in the qualitative calculus which represent human observable events, e.g. following and overtaking. The system can then be used to recognise subsequent selected event occurrences or unusual behaviours. Although not explicitly encoded as abductive reasoning, this recognition phase can be viewed in this way: at each time step various behaviours may be possibly applicable given the current observations and the system keeps track of all the different possible explanations of the data. The system actually has recorded statistical frequency data during the learning phase and could use this to rank order the hypotheses. In the actual implementation they are all equally ranked and kept as possible explanations until subsequent observations rule out particular behaviours.

In newer work [36] based on [37, 38], we are also interested in automatically inferring models of object interactions that can be used to interpret observed behaviour within a scene. Low-level computer vision techniques together with an attentional control mechanism are used to identify interesting incidents or events that occur in the scene over long periods of time. A data driven approach has been taken in order to automatically infer discrete and abstract representations (*symbols*) of primitive object interactions; this can be viewed as learning the basic qualitative spatial language of  $\Sigma$ . These symbols are then used as an alphabet to infer the high level structure of typical interactive behaviour using variable length Markov models (VLMs) [39, 40]. VLMs deal with a class of random processes in which the memory length varies, in contrast to an  $n$ th-order Markov model for which the memory length is fixed. They have been previously used on the data compression [41, 42] and language modelling domains [39, 40, 43] and recently, they have been successfully introduced in the computer vision domain for automatically inferring stochastic models of the high level structure

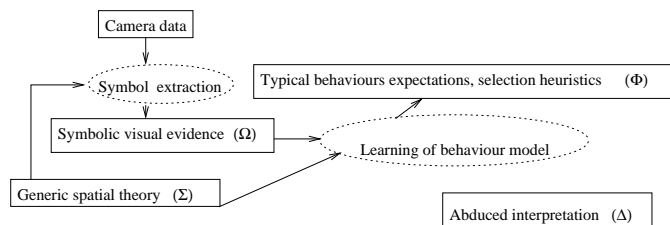
of complex and semantically rich human activities [37, 38]. It should be noted that although we are currently concentrating on applications within the traffic domain, our method is applicable to the general automatic surveillance task since it does not assume *a priori* knowledge of a specific domain. It is also worth pointing out explicitly that the use of probabilistic reasoning here contrasts with conventional low level use of probabilities – here the Markov model concerns high level semantic notions.



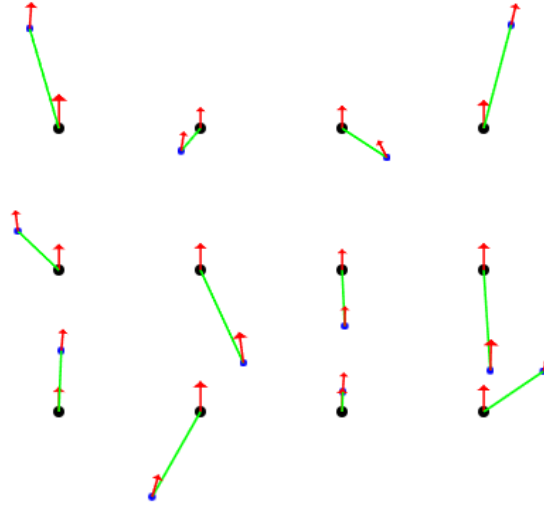
**Fig. 2.** System Overview: abducting interpretations



**Fig. 3.** System Overview: learning spatial relations



**Fig. 4.** System Overview: learning behavioural models



**Fig. 5.** Learnt primitive interactions – traffic domain example. The two dots represent pairs of close vehicles (distinguished by the size of the dot). The arrows show their direction of movement and the connecting vector their relative orientation. These patterns represent typical “midpoints” as result of clustering the input data into  $n$  different conceptual “regions”. Note how the 12 relations naturally cluster into virtually symmetric pairs, e.g. the middle two prototypes on the first line.

Figures 2, 3 and 4 give an overview of the system including the learning element of the architecture whereby the typical behaviour patterns are learned, stored and used to drive interpretation. These are the behavioural patterns,  $\mathcal{P}$ , referred to in section 4 above. A real time computer vision system [44] detects and tracks moving objects within a scene. For each moving object scene feature descriptors are extracted that describe its relative motion and spatial relationship to all moving objects that fall within its attentional window (see [36] for details). These scene feature descriptors are invariant of the absolute position and direction of the interacting objects within a scene and are termed  $\Omega$  in the figures.

Figure 2 shows the operation of the system once learning has taken place. Figure 3 shows how learning of part of the generic spatial theory  $\Sigma$  might take place. Figure 5 illustrates learnt primitive interactions for a traffic domain example application [36]. These can be viewed as a qualitative discretisation of the continuous relational space. Whereas in a conventional QSR representation, the discretisation would be manually preassigned, we are able to learn a representation which maximises the discernability given a granularity (i.e. the number of relations desired). The system can currently be used to recognize typical interactive behaviour within the traffic domain and identify atypical events. These

learnt primitive interactions effectively form part of the background spatial theory, which is labelled  $\Sigma_2$  in figure 3.

In figure 4 we show how behaviours might be learnt. In [36] we use a statistical learning framework [45] where discrete representations of interactive behaviours can be learned by modelling the probability distribution of the primitive interactions. VLMs are then used to efficiently encode the sequences of these learned primitive patterns corresponding to observed interactive behaviour.

## 6 Incorporating QSR into the Heart of a Computer Vision System

As already noted, QSR has been used as a post-processing method on the output of a quantitative real-world analysis system (such as a vision system), e.g. [4]. In this section we will propose an alternative approach that puts QSR methods at the heart of a computer vision system. This will have the effect of constraining the output of the system to be logically consistent (with respect to the QSR theory embodied in  $\Sigma$  and  $\Phi$ ). This is done by using low level CV algorithms (e.g. colour region segmentation algorithms) that draw no conclusions about the nature or structure of the data. The output of the low level process thus makes as few semantic inferences as possible. For example it will not embody an “object tracker” as that would presuppose the ability to recognise objects. What it does do, is at each time step (frame) to distinguish certain spatial elements, and assign certain qualitative properties to them (such as colour, texture and qualitative spatial relationships). The sequence of these outputs comprises  $\Omega$ . The spatial elements in  $\Omega$  thus become the primitive spatial elements (rather than the original pixels); like pixels they may be “mixed”, in the sense that they contain elements of different objects, but they will never be split apart, but the heterogeneity will be symbolically reasoned about.

The higher level reasoning component then comprises three principal mechanisms:

- A qualitative spatio-temporal reasoner which uses  $\Sigma$  and performs certain inferences such as checking consistency both statically and with respect to continuous motion for example.
- An abduction engine which uses  $\Phi$  in conjunction with the other data to generate hypotheses, i.e. possible (partial) explanations.
- A technique for handling uncertainty: the behaviours in  $\Phi$  may be probabilistic or have other metadata indicating their absolute or relative likelihood. This component should ensure that the most likely explanation is chosen.

The objective of our proposed system is to explain a complete observation (or set of observations) rather than a subset of the objects within the scene as is traditionally the task. This removes the logical distinction between objects and ‘background’ (the rest of the scene). This has the practical disadvantage of causing the computational cost of any QSR method to be exponential in the granularity of the problem.

To cope with the computational explosion caused by this explicit use of symbolic spatial reasoning, we borrow two techniques used in human perception: attention and multi-resolution/scale processing. A good attention mechanism commonly used in CV is motion. This is highly suitable for online reasoning as moving areas contain more spatio-temporal object information than static areas and as such require processing at a higher rate/resolution.

### 6.1 Scene Understanding as Scene Explanation

We wish to represent a single observation of a scene as resulting from a number of objects (with no concept of background). In the real world, scene description is not this simple as objects exist in a conceptual hierarchy. Figure 6 gives an example of such a hierarchy.

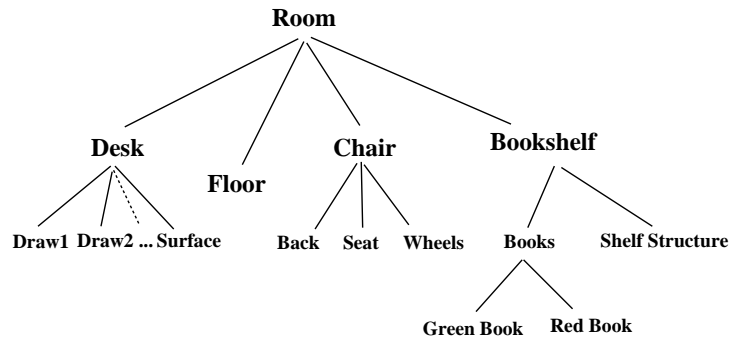


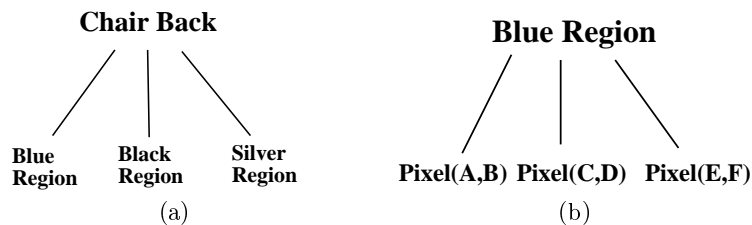
Fig. 6. Conceptual Hierarchy of Objects for a Room

Figure 6 illustrates that many conceptual objects are in fact composite objects constructed by the combination of simpler objects or composite objects. This conceptual hierarchy sits on a single level observational hierarchy in which base level conceptual objects are divided into non-semantic observational regions. An example of this is given in figure 7(a).

From the point of view of a CV system (or human vision system) the conceptual hierarchy sits on a sensory hierarchy with atomic sensory components (pixels in the case of a CV system). An example of this is given in figure 7(b).

### 6.2 Automatic Building of Object Hierarchies

The purpose of any scene understanding system is to automatically build complete or (more usually) partial object hierarchies of the nature of those described in the previous section from the bottom up. Alternatively, *a priori* information may be used in a 'hypothesise and test way' to build object hierarchies from the top down. The bottom up approach is essentially limited to building sensory and



**Fig. 7.** (a) Observational Hierarchy of Chair Object (b) Sensory Hierarchy of a Region: a particular region (“Blue Region”) is composed of a number of pixels at particular  $x, y$  coordinates.

observational hierarchies as no *a priori* conceptual information is available. The top down approach can build complete hierarchies; however an *a priori* model for any object or composite object that may occur in a scene must be available. This is a problem if we wish to build complete hierarchies of complex scenes. In many real world CV systems a combination of the top down and bottom up approaches is used [44, 46]; however the interface of these two approaches is often *ad hoc*. This can lead to errors and logical inconsistencies in the final scene analysis. We propose to use QSR to interface low level (bottom up) approaches with high level (top down) methods in such way that low level logical inconsistencies do not occur in the high level interpretation. An example of this would be to use continuity networks such as the one in figure 1 to filter out low level spatio-temporal data which are not continuous with respect to this diagram (e.g. if disconnected regions are immediately afterwards partially overlapping). A refinement to this approach is to use continuity networks which are specialised to the kinds of objects involved. In [47, 30] we distinguish various weaker notions of continuity which may be appropriate for certain kinds of objects and correspondingly weaker conceptual neighbourhood diagrams. If the vision system can recognise the types of the objects involved, then the notion of continuity can be correspondingly specialised.

We propose to use bottom up CV to build sensory hierarchies that describe the entire image and use abduction to generate (over time) a set of logically consistent hypotheses for the complete scene description hierarchy. Higher level (top down) CV methods incorporating *a priori* knowledge would then be used to validate and rank these hypothesis and assign semantic labels. Objects in a scene for which no *a priori* information exists would remain unvalidated and unlabelled; however this would be explicitly flagged by the system and could be used as the basis of a novel object learning system.

Past time hypotheses may be declared invalid given subsequent observations and deleted. For reasons of computational tractability it may be necessary to only consider a small window into the past when evaluating the validity of past hypotheses.

Our approach to abduction is outlined in more detail in [48]. In essence, the problem is to determine, given a background spatio-temporal theory, a set of typical patterns of behaviour and a set of qualitative spatio-temporal observations, what actual objects and behaviours could explain the observations.

### 6.3 Reasoning over Time and Hypothesis Verification

QSR and abduction will generate scene hypotheses for the complete scene description hierarchy at the current timestep and at previous timesteps and the validity relationships between these over time. This is illustrated in figure 8. As can be seen, in general there will be more than one possible explanation abduced and a way of rank ordering the various hypotheses offered as explanations will be needed too. In [48] we give some logic based techniques whereby a preferred hypothesis (or set of preferred hypotheses) might be selected. In the dynamic case we are considering here, we will want to carry forward multiple hypotheses from one frame to the next and use information gained from future frames as well as statistically and a priori knowledge based heuristics to choose a single preferred hypothesis when required.

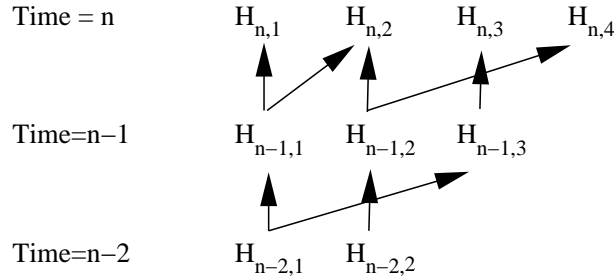


Fig. 8. Hypotheses are Generated over Time by QSR and Abduction

Using *a priori* knowledge, a hypothesis may be assigned low probability at the current timestep and this can be propagated as a future hypothesis. *A priori* knowledge can exist on many levels from detailed object models (e.g. 3D shape or texture models) to more general object class models (e.g walls are static and of homogeneous colour and texture). In a hypothesis verification scenario these different types of *a priori* knowledge may be combined so as to verify the complete hypotheses.

Methods used to encapsulate typical (statistical) *a priori* spatial or temporal object information in the scheme described so far must be able to take as their input an object (or composite object) description hierarchy (or sequence of these). These hierarchies may be thought of as a list of atomic elements containing their properties (in addition to structural information which may be ignored if necessarily). These atomic elements may be at the sensory level (pixels) or the

observational level (homogeneous regions). This does not fit well with the traditional N-dimensional object models described in section 2. These models must be adapted or replaced to fit in with our proposed variable length list object descriptions. How such methods are to be formulated is the subject of current work and beyond the scope of this paper; however it should be noted that comparisons between model and observation are on an object-object basis rather than an object-scene basis as with many traditional CV methods. This allows observation to model matching in addition to model to observation matching.

## 7 Future Work

We are planning a wide variety of future work in order to flesh out and validate our proposed architecture, not only within the traffic domain but also in another domain, for example a “kitchen or table top scenario”. There is theoretical work to do as well as actually implementing a system conforming to the ideas presented here.

In particular, further research in qualitative spatial and spatio-temporal representation and reasoning will be required. Much work has concentrated on topological and mereotopological calculi to date as indicated in [12]. New calculi such as the occlusion calculus [49], are being specifically developed for cognitive vision – the occlusion calculus being specifically targeted at the problem of reasoning about the topology of occluded regions.

However, in order to model and distinguish between different kinds of objects that might be found in visual scenes, not only will other aspects, such as orientation and size become important, but more particularly, the notion of qualitative shape will need to be further developed so as to be useful for cognitive vision. There is already some work on orientation (e.g. [18, 50]), though this is largely point based, rather than region based, which may prove to be more useful. There is relatively little work still on qualitative shape representations. Existing work includes boundary based approaches [51, 52], representation through elongation and symmetric aspects [53], and the use of a convex hull primitive (which essentially gives an affine geometry [54]). However the utility of these approaches applied to cognitive vision has not been tested to any great extent.

It is worth pointing out that although in general it is very hard to represent shape in a qualitative way since very small changes in shape may lead to very different functionality (e.g. consider interlocking gears), in cognitive vision the task is not so much to reason about kinematics (or similar predictive/analytical tasks which require detailed shape knowledge) but rather simply to categorise object shape in order to classify and recognise different kinds of objects. Arguably this task will be easier, but this has not been particularly investigated.

A vital aspect of a cognitive vision architecture is the ability to represent and reason about extended event sequences. Although VLMMs have been successfully applied in computer vision in order to represent long-term behaviours, criticisms may be made of them that they have no semantics in themselves. The use of a qualitative spatio-temporal representation within a logical framework

holds out the promise of a formally defined semantics, and a richer vocabulary to describe extended behaviours. Moreover, the notion of continuity present in the conceptual neighbourhoods of a qualitative spatial calculus may be used to help constrain the learning and interpretation of event sequences. However more research is required in order to validate this hypothesis and to develop a qualitative spatio-temporal theory that is well adapted to the demands of cognitive vision.

Finding control architectures to moderate the inferential mechanism to process data efficiently is a key research question, for example to develop a controllable attentional mechanism. Also on our priority list is to develop further techniques for learning the background knowledge  $\Sigma$  and particularly  $\Phi$ .

We also plan to consider other input features apart from orientation, relative direction of motion and distance, e.g. acceleration and shape. Another consideration is to consider how non interactive behaviours may be represented and reasoned about (at present the feature descriptions are always between pairs of moving objects). Finally, we hope to integrate our approaches with some of those in our partners in the CogVis project IST-2000-29375 and to take account of other work such as the approach to learning from low level data of [55].

Finally, we plan to consider the evaluation of our cognitive vision system: under what circumstances would we say that we have succeeded? Clearly, we can inspect the internal architecture of the system and the extent to which it has high level representations, the extent to which it can learn and meet the other considerations mentioned in the introduction. A further criterium would be to evaluate with respect to human visual cognition; for example Tversky [56, 57] has investigated the perception of the event structure of a video sequence by human subjects – can we produce a cognitive vision system which can infer a similar structure?

## Acknowledgements

The support of the EPSRC under grant GR/M56807 and the EU under IST-2000-29375 is gratefully acknowledged.

## References

1. Yanai, K., Deguchi, K.: Recognition of indoor images employing qualitative model fitting and supporting relation between objects. In Sanfeliu, A., Villanueva, J., Vanrell, M., Alquezar, R., Eklundh, J.O., Aloimonos, Y., eds.: Proceedings 15th International Conference on Pattern Recognition. Volume 1., Barcelona, Spain, IEEE Press (2000) 964–967
2. Howarth, R.: Interpreting a dynamic and uncertain world: High-level vision. *Artificial Intelligence Review* **9** (1995) 37–63
3. Buxton, H., Howarth, R.: Spatial and temporal reasoning in the generation of dynamic scene descriptions. In Rodríguez, R.V., ed.: Proceedings on Spatial and Temporal Reasoning, Montréal, Canada, IJCAI-95 Workshop (1995) 107–115

4. Fernyhough, J., Cohn, A., Hogg, D.: Constructing qualitative event models automatically from video input. *Image and Vision Computing* **18** (2000) 81–103
5. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Training models of shape from sets of examples. In: *Proc. British Machine Vision Conference*. (1992) 9–18
6. Baumberg, A., Hogg, D.: Learning flexible models from image sequences. In: *European Conference on Computer Vision*, Springer Verlag (1994) 299–308
7. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. In: *Proc. First International Conference on Computer Vision*. (1989) 259–268
8. Blake, A., Curwen, R., Zisserman, A.: A framework for spatiotemporal control in the tracking of visual contours. *International Journal of Computer Vision* **11** (1993) 127–145
9. Turk, M., Pentland, A.: Eigenfaces for recognition. *Journal of Cognitive Neuroscience* **3** (1991) 71–86
10. Rabiner, L.: A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77** (1989) 257–286
11. Starner, T., Pentland, A.: Real-time american sign language recognition from video using hidden markov models. In: *Int. Symposium on Computer Vision*. (1995)
12. Cohn, A.G., Hazarika, S.M.: Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae* **46** (2001) 1–29
13. Cohn, A.G., Bennett, B., Gooday, J., Gotts, N.: RCC: a calculus for region based qualitative spatial reasoning. *GeoInformatica* **1** (1997) 275–316
14. Egenhofer, M., Franzosa, R.: Point-set topological spatial relations. *International Journal of Geographical Information Systems* **5** (1991) 161–174
15. Cohn, A.G., Gotts, N.M.: Representing spatial vagueness: a mereological approach. In L C Aiello, J.D., Shapiro, S., eds.: *Proceedings of the 5th conference on principles of knowledge representation and reasoning (KR-96)*, Morgan Kaufmann (1996) 230–241
16. Clementini, E., Di Felice, P.: Approximate topological relations. *International Journal of Approximate Reasoning* **16** (1997) 173–204
17. Schlieder, C.: Reasoning about ordering. In A Frank, W.K., ed.: *Spatial Information Theory: a theoretical basis for GIS*. Number 988 in *Lecture Notes in Computer Science*, Berlin, Springer Verlag (1995) 341–349
18. Isli, A., Cohn, A.: A new approach to cyclic ordering of 2d orientations using ternary relation algebras. *Artificial Intelligence* **122** (2000) 137–187
19. Frank, A.U.: Qualitative spatial reasoning about distance and directions in geographic space. *Journal of Visual Languages and Computing* **3** (1992) 343–373
20. Meathrel, R.C., Galton, A.P.: A heirarchy of boundary-based shape descriptors. In Nebel, B., ed.: *Proc. 17th IJCAI*, Morgan Kaufmann (2001) 1359 – 1364
21. Bennett, B., Cohn, A.G., Torrini, P., Hazarika, S.M.: Describing rigid body motions in a qualitative theory of spatial regions. In Kautz, H.A., Porter, B., eds.: *Proceedings of AAAI-2000*. (2000) 503–509
22. Cristani, M., Cohn, A., Bennett, B.: Spatial locations via morpho-mereology. In: *Proc. KR'2000*, Morgan Kaufmann (2000)
23. Cohn, A.G., Bennett, B., Gooday, J., Gotts, N.: Representing and reasoning with qualitative spatial relations about regions. In Stock, O., ed.: *Temporal and spatial reasoning*, Kluwer (1997)
24. Wolter, F., Zakharyashev, M.: Spatio-temporal representation and reasoning based on RCC-8. In: *Proceedings of the seventh Conference on Principles of Knowledge Representation and Reasoning*, Morgan Kaufman (2000) 3–14

25. Wolter, F., Zakharyashev, M.: Qualitative spatio-temporal representation and reasoning: a computational perspective. In: Exploring Artificial Intelligence in the New Millennium. Morgan Kaufmann (To appear)
26. Bennett, B., Cohn, A., Wolter, F., Zakharyashev, M.: Multi-dimensional modal logic as a framework for spatio-temporal reasoning. *Applied Intelligence* (2002) To appear.
27. Hayes, P.J.: Naive physics I: Ontology for liquids. In Hobbs, J.R., Moore, B., eds.: *Formal Theories of the Commonsense World*. Ablex (1985) 71–89
28. Muller, P.: A qualitative theory of motion based on spatio-temporal primitives. In Cohn, A.G., Schubert, L.K., Shapiro, S., eds.: *Principles of Knowledge Representation and Reasoning: Proceedings of the 6th International Conference (KR-98)*, Morgan Kaufman (1998) 131–141
29. Muller, P.: Space-time as a primitive for space and motion. In Guarino, N., ed.: *Formal ontology in information systems: Proceedings of the 1st international conference (FOIS-98)*. Volume 46 of *Frontiers in Artificial Intelligence and Applications*, Trento, Italy, Ios Press (1998) 63–76
30. Hazarika, S.M., Cohn, A.G.: Qualitative spatio-temporal continuity. In Montello, D.R., ed.: *Spatial Information Theory: Foundations of Geographic Information Science; Proceedings of COSIT'01*. Volume 2205 of LNCS., Morro Bay, CA, Springer (2001) 92–107
31. Cui, Z., Cohn, A.G., Randell, D.A.: Qualitative simulation based on a logical formalism of space and time. In: *Proceedings of AAAI-92*, Menlo Park, California, AAAI Press (1992) 679–684
32. Shanahan, M.: Noise, non-determinism and spatial uncertainty. In: *Proceedings of AAAI-97*. (1997) 153–158
33. Shanahan, M.: A logical account of the common sense informatic situation for a mobile robot. *Electronic Transactions on Artificial Intelligence* (1999)
34. Remolina, E., Kuipers, B.: A logical account of causal and topological maps. In: *Proceedings of Seventeenth International Conference on Artificial Intelligence (IJCAI-01)*. Volume I, Seattle, Washington, USA (2001) 5–11
35. Shanahan, M.: A logical account of perception incorporating feedback and expectation. In: *Proc. 8th Int. Conf. on Knowledge Representation and Reasoning*, San Mateo, Morgan Kaufmann (2002)
36. Galata, A., Cohn, A.G., Magee, D., Hogg, D.: Modelling interaction using learnt qualitative spatio-temporal relations and variable length markov models. In: *Proc. European Conference on AI (ECAI)*. (2002)
37. Galata, A., Johnson, N., Hogg, D.: Learning behaviour models of human activities. In: *British Machine Vision Conference, BMVC'99*. (1999)
38. Galata, A., Johnson, N., Hogg, D.: Learning Variable Length Markov Models of Behaviour. *Computer Vision and Image Understanding (CVIU) Journal* **81** (2001) 398–413
39. Ron, D., Singer, S., Tishby, N.: The Power of Amnesia. In: *Advances in Neural Information Processing Systems*. Volume 6. Morgan Kauffmann (1994) 176–183
40. Guyon, I., Pereira, F.: Design of a Linguistic Postprocessor using Variable Memory Length Markov Models. In: *International Conference on Document Analysis and Recognition*. (1995) 454–457
41. Cormack, G., Horspool, R.: Data Compression using Dynamic Markov Modelling. *Computer Journal* **30** (1987) 541–550
42. Bell, T., Cleary, J., Witten, I.: *Text Compression*. Prentice Hall (1990)
43. Hu, J., Turin, W., Brown, M.: Language Modelling using Stochastic Automata with Variable Length Contexts. *Computer Speech and Language* **11** (1997) 1–16

44. Magee, D.: Tracking multiple vehicles using foreground, background and motion models. In: Proc. ECCV Workshop on Statistical Methods in Video Processing. (2002)
45. Johnson, N., Hogg, D.: Learning the Distribution of Object Trajectories for Event Recognition. *Image and Vision Computing* **14** (1996) 609–615
46. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfnder: Real-time tracking of the human body. *IEEE Transactions on PAMI* **19(7)** (1997) 780–785
47. Cohn, A.G., Hazarika, S.M.: Continuous transitions in mereotopology. In: Commonsense-2001: 5th Symposium on Logical Formalizations of Commonsense Reasoning. (2001)
48. Hazarika, S.M., Cohn, A.G.: Abducing qualitative spatio-temporal histories from partial observations. In: Proc. 8th Int. Conf. on Knowledge Representation and Reasoning, San Mateo, Morgan Kaufmann (2002)
49. Randell, D., Witkowski, M., Shanahan, M.: From images to bodies: Modelling and exploiting spatial occlusion and motion parallax. In: Proc. IJCAI, Morgan Kaufmann (2001)
50. Freksa, C.: Using orientation information for qualitative spatial reasoning. In Frank, A.U., Campari, I., Formentini, U., eds.: Proc. Int. Conf. on Theories and Methods of Spatio-Temporal Reasoning in Geographic Space, Berlin, Springer-Verlag (1992)
51. Meathrel, R.C., Galton, A.: A hierarchy of boundary-based shape descriptors. In: Proc. IJCAI. (2001) 1359–1364
52. Jungert, E.: Symbolic spatial reasoning on object shapes for qualitative matching. In Frank, A.U., Campari, L., eds.: *Spatial Information Theory: A Theoretical Basis for GIS*. Lecture Notes in Computer Science No. 716, COSIT'93, Springer-Verlag (1993) 444–462
53. Clementini, E., Di Felice, P.: A global framework for qualitative shape description. *Geoinformatica* **1** (1997) 1–17
54. Davis, E., Gotts, N.M., Cohn, A.G.: Constraint networks of topological relations and convexity. *Constraints* **4** (1999) 241–280
55. Kaelbling, L.P., Oates, T., Hernandez, N., Finney, S.: Learning in worlds with objects. In Cohen, P.R., Oates, T., eds.: *Learning Grounded Representations*. Number Technical Report SS-01-05, AAAI Press (2001) 31–36
56. Zacks, J., Tversky, B., Iyer, G.: Perceiving, remembering and communicating structure in events. *Journal of Experimental Psychology: General* **136** (2001) 29–58
57. Zacks, J., Tversky, B.: Event structure in perception and conception. *Psychological Bulletin* **127** (2001) 3–21