Combining Logic and Probability in Tracking and Scene Interpretation

Brandon Bennett

University of Leeds, School of Computing
Leeds, LS2 9JT, UK
brandon@comp.leeds.ac.uk

Abstract. The paper gives a high-level overview of some ways in which logical representations and reasoning can be used in computer vision applications, such as tracking and scene interpretation. The combination of logical and statistical approaches is also considered.

Keywords. Vision, Tracking, Logic, Probability, Spatio-Temporal Continuity

1 Introduction

Computer Vision was among the first problems identified as a goal of Artificial Intelligence. In these early days, logical reasoning was seen as a general purpose mechanism for modelling intelligent behaviour. Hence, research in Computer Vision often involved development of semantic representations and inference mechanisms (see e.g. the papers in the collection [1]). Indeed the development of reasoning techniques such as constraint propagation is closely associated with early work in computer vision (*vis* [2]).

Since these early days there has been a marked divergence between the fields of Knowledge Representation and Computer Vision. Whereas, Knowledge Representation still uses logical representations and reasoning as one of its primary tools, Computer Vision research has moved away from this, embracing statistical techniques as the underpinning for most of its algorithms. Perhaps the main reason for this is that many problems in visual processing turned out to be much harder than had originally been anticipated. Interpretation of visual scenes tends to be highly unreliable (except in highly constrained artificial setting). Hence, statistics are employed to find the most likely interpretations out of many possibilities.

However, it is now becoming apparent that statistical methods alone have limitations. In particular, it is very difficult to combine localised probabilistic information into a coherent overall description without taking account semantic constraints between separate pieces of information. Logical reasoning can provide a powerful mechanism for determining consistent possibilities. Also, by describing the conceptual structure of possible situations, semantic knowledge may be used to guide the search for plausible interpretations of incomplete or ambiguous data.

Dagstuhl Seminar Proceedings 08091
Logic and Probability for Scene Interpretation
http://drops.dagstuhl.de/opus/volltexte/2008/1612

2 Enhancing Tracking by Enforcing Spatio-Temporal Continuity

As an illustration of how logic may be used in visual interpretation, I first consider a problem of object tracking and recognition, which was tackled at Leeds, as part of the European *Cognitive Vision* project.[1] This work involved collaboration between the vision research group, whose work has resulted in a number of tracking and object recognition systems, and the Knowledge representation and Reasoning group, whose focus has been particularly on spatial and spatio-temporal representation and reasoning [3,4,5].

The goal of tracking is to extract information about the positions of moving objects from a dynamic visual scene. A common approach is to first identify moving objects by use of so-called 'blob-tracker', which uses techniques such as background subtraction [6,7,8]. Once areas of an image have been identified as containing a moving object, particular objects within these areas can be identified by the same methods as for static images.

With real video data, such an approach will normally give very poor results. Because the objects are moving, they will be seen from many different angles. Moreover occlusions may occur, where one object overlaps or perhaps completely hides another. Under such conditions, it is not surprising that recognition algorithms developed for static images are extremely unreliable.

But it is evident that by separating the tracking and recognition tasks, much useful information concerning the objects has been thrown away. In particular, the constraints of basic physics mean that the tracked objects must move *continuously* in time and space. Thus a possible labelling of the objects corresponding to the moving blobs detected by the tracker must be consistent with this continuity constraint. Handling such constraints in terms of joint probabilities would require very sophisticated statistical modelling and intensive computation. However, within a logical framework, such constraints are relatively easy to state and reason with.

The tracker used in this research generates a sequence of boxes indicating the approximate locations of moving objects. This output can be simplified so

[1] COGVIS: http://www.comp.leeds.ac.uk/vision/cogvis/.

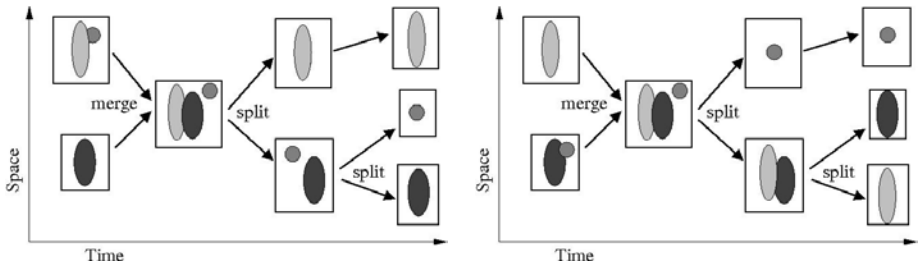

Fig. 1. Two possible object labellings of tracker box output.

that frame sequences over which there is no change in the number of boxes are treated as a single state. The result is a series of states where between successive states is a *merge* or *split* event. These occur when objects move together (possibly occluding each other) and when clustered objects separate, as illustrated in Figure 1. This representation allows continuity to be captured by very simple constraints: when two boxes merge, the objects that were in those boxes must be present in the merged box; and, when a box splits, the objects in the original box must be distributed between the two resulting boxes.

Figure 1 shows how an output of this kind can be given different labellings consistent with spatio-temporal continuity. Since the logical constraints cannot decide between these, it is here that statistical information can be used to choose the most likely labelling. Whereas both the statistical and logical methods used separately cannot decide reliably between different interpretations, our experimental results indicate that their combination greatly improves performance. By taking the logically consistent model with the best statistical support, we can achieve very high accuracy. Moreover, enforcing spatio-temporal consistency enables us to give correct labellings even when objects are occluded or completely hidden.

More details of how logical and statistical knowledge are combined and the results obtained are given in the following publications: [9,10,11].

3 Bi-Directionality of Visual Information Processing

One of the primary goals of machine vision is to extract high level descriptions from low level data. Partial success has been achieved by methods that are either:

- *Bottom-Up* — where high level entities/relationships are synthesised from low level data.
- *Top-Down* — where high level models are used to guide the search for evidence in low level data.

Improved performance is often achieved by systems that employ a *combination* of bottom up and top down mechanisms. But it seems to be increasingly apparent (especially in much of the work reported at this Dagstuhl workshop) that visual interpretation requires information processing in a style that is *intrinsically bi-directional*.

By its nature, logic is undirected in the way it manipulates information. Axioms may involve concepts and relations at different levels of abstraction. For example, consider the following formula, which characterises the high-level concept Pillar in terms of low-level geometrical properties:

D1) $\forall x[\mathsf{Pillar}(x) \leftrightarrow (\mathsf{Solid}(x) \wedge \mathsf{Cylinder}(x) \wedge \mathsf{Vertical}(\mathsf{princ\_axis}(x)))]$

This equivalence could be exploited in either direction: we may have reason to believe a pillar is present (perhaps because of an even higher level description of a building) and use it to guide the search for this object within the geometry

of a scene; or we may have an arbitrary scene which we process in terms of its low-level geometrical structure and then use definitions such as this to identify high-level objects.

More generally, the pillar definition could function in both directions within a single scene interpretation algorithm. To illustrate this, I first give a further axiom defining (roughly) the concept of Arch as consisting of two (or more) pillars supporting a lintel.

$$\textbf{D2)}\ \mathsf{Arch}(x) \leftrightarrow \exists yzw[\neg(x=y) \wedge \mathsf{Part}(y,x) \wedge \mathsf{Part}(z,x) \wedge \mathsf{Part}(w,x) \wedge \mathsf{Pillar}(y) \wedge \mathsf{Pillar}(z) \wedge \mathsf{Lintel}(w) \wedge \mathsf{Supports}(y,w) \wedge \mathsf{Supports}(z,w)]$$

Given these logical definitions, an algorithm could work as follows. First it uses definitions such as **D1** to look for basic objects like pillars based on their geometrical definitions. At this stage one would probably require a relatively high degree of conformity to the geometrical specification to avoid excessive generation of objects that may not actually be present. Then using definitions such as **D2** the possible existence of higher level composite objects such as an arch can be inferred, even if certain components of these composites have not been identified. In this case, having found one pillar, one can infer that an arch may be present if a second pillar and a lintel can be found in a suitable geometric configuration. The geometrical definitions of pillar and arch could then be re-applied to try to find these components to complete the arch. At this stage, a weaker level of geometrical conformity could be required, since, having previously identified a pillar, one already has some evidence of the existence of an arch.

General purpose logical reasoners do not constrain the directionality of information manipulation and hence may provide a natural vehicle for implementing bi-directional visual processing architectures. However, the non-directedness of general proof procedures quickly leads to *intractability*, so it is likely that a more task-oriented control mechanism would be required to achieve reasonable performance. One way in which logical theories might be made more amenable to the kinds of reasoning relevant to scene interpretation is to emphasise *definitional* structure within the logical theory. Definitional axioms make explicit two of the most significant directions for inference in scene interpretation: inferring instances of higher level concepts from information expressed in terms of lower-level concepts; and conversely, inference from the hypothesised existence of instances of high-level concepts to requirements on low-level information which can support the hypothesis. Distinguishing the lowest level *primitives* from defined concepts also facilitates grounding and model-based reasoning techniques.

4 Issues in Logical Modelling

We have seen that certain semantic constraints relevant to tracking can be captured by relatively simple logical rules. But scene interpretation in general involves an extremely rich semantics and to cover all aspects would require a complex and extensive theory. Moreover this domain involves many subtleties which present significant challenges for logical modelling.

One aspect of visual scenes that must be taken into account is the ontological distinction between actual objects and their appearances. Most formal theories of the physical world directly represent objects and their attributes, implicitly assuming that the domain of objects and the conditions under which they possess these attributes are clear. But in most visual scene analysis problems these assumptions cannot be made. On the one hand, the domain of objects may be unknown or open-ended, and on the other we do not have completely reliable methods for detecting attributes of objects.

Although the relation between objects and appearances is a major topic in philosophical meta-physics it is seldom covered in modern logic or AI representations. The following papers may provide a staring point for developing a suitable logical formalism: [12,13].

Another important issue is the relationship between the logical theory and information of a statistical nature. In the tracking work described above a rather simple approach was used, we identified a model that was consistent with the logical theory and best supported by the statistical information. To evaluate the statistical support we simply summed over a number of localised probability measures. Although this gave good results, it lacked any rigorous theoretical basis.

In order to develop a more principled approach, we need to employ a much more sophisticated logical treatment. One possibility is to use a fully fledged logic of probability such as that proposed by Halpern [14]. Use of such a logic would make explicit certain aspects of probability which are often glossed over in its application to computer vision applications. One such aspect is the distinction between various ways in which probabilities may be applied to propositional statements. Halpern identifies the following distinct modes in which logical statements may be given probabilities:

- $\mathsf{Prob}(\phi)$ — probability of ϕ relative to a probability distribution over possible worlds.
- $\mathsf{Prob}_x(\phi(x))$ — probability of $\phi(x)$ relative to a probability distribution over the domain of individuals.

In the context of scene interpretation, these two types of probability are often compounded together in subtle ways. For instance, a particular type of object, T, may have a certain probability p of possessing an attribute ϕ; and a given feature detection attribute may have a certain probability q of detecting ϕ if present. The first figure p is is a probability distribution over individuals of the domain, whereas q is a probability relative to the distribution of possible worlds. Hence, the overall probability that an object of type T will be detected as possessing attribute ϕ involves a complex combination of statistics evaluated over the domains of both individuals and possible worlds.

In the context of reasoning about actions and changes a rich theory of how to reason about noisy sensor has been proposed by Bacchus *et al.* [15]. With some modification such a theory might also be applied as a basis for integrating logic and probability. However, due to the complexity of reasoning with formal theories

and the large amount of data that vision systems must deal with, it is likely that considerable simplification would need to be made to achieve effectiveness in real vision applications.

5 Conclusion

The aim of this short article has been to indicate some promising ways in which logical reasoning and probabilistic information can be combined in order to support algorithms for visual scene interpretation. The method of finding a logically consistent model that is best supported by statistical information has been found to be very effective in particular experiments on an object tracking problem. But the basic idea is extremely general and could be applied in many areas of computer vision.

I have suggested that logic can provide a representation suitable for specifying and guiding bi-directional processing of visual information, in which bottom up construction of complex objects from their constituents can be flexibly combined with top down search for the constituents of an object whose existence is hypothesised. I also pointed out that the use of logic in computer vision applications lacks theoretical foundations both in respect of formalising an ontology of objects and appearances and in respect to formalising the semantics of the attribution of probabilities to propositions.

In summary, the need to use logical reasoning in scene interpretation is becoming increasingly apparent, and there are many promising approaches that could be taken. However, there are considerable gaps in the theory of how logic and probability should be combined, and logical theories suitable for use in visual processing applications are at an early stage of development.

References

1. Winston, P., ed.: The Psychology of Computer Vision. McGraw-Hill (1975)
2. Waltz, D.L.: Understanding line drawings of scenes with shadows. In Winston, P., ed.: The Psychology of Computer Vision. McGraw-Hill (1975) 19–92
3. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connection. In: Proc. 3rd Int. Conf. on Knowledge Representation and Reasoning, San Mateo, Morgan Kaufmann (1992) 165–176
4. Bennett, B.: Space, time, matter and things. In Welty, C., Smith, B., eds.: Proceedings of the 2nd international conference on Formal Ontology in Information Systems (FOIS'01), Ogunquit, ACM (2001) 105–116 http://www.comp.leeds.ac.uk/qsr/pub/bennett-fois01.pdf.
5. Hazarika, S.M., Cohn, A.G.: Abducing qualitative spatio-temporal histories from partial observatio ns. In Fensel, D., Guinchiglia, F., McGuinness, D., Williams, M.A., eds.: Proceedings of the Eight Conference on Principles of Knowledge Representation and Reasoning (KR 2002), Morgan Kaufmann (2002) 14–25
6. Haritaoglu, I., Harwood, D., Davis, L.: W4: Who? when? where? what? a real time system for detecting and tracking people. In: Proc. International Conference on Automatic Face and Gesture Recognition. (1998) 222–227

7. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: Proc. Computer Vision and Pattern Recognition. (1999) 246–252
8. Magee, D.R.: Tracking multiple vehicles using foreground, background and motion models. Image and Vision Computing **20(8)** (2004) 581–594
9. Bennett, B., Magee, D.R., Cohn, A.G., Hogg, D.C.: Using spatio-temporal continuity constraints to enhance visual tracking of moving objects. In Saitta, L., ed.: Proceedings of the 16th European Conference on Artificial Intelligence (ECAI-04), ECCAI, IOS Press (2004) 922–926
10. Bennett, B., Cohn, A.G., Magee, D.: Enforcing global spatio-temporal consistency to enhance reliability of moving object tracking and classification. KI **19** (2005) 32–35
11. Bennett, B., Magee, D.R., Cohn, A.G., Hogg, D.C.: Enhanced tracking and recognition of moving objects by reasoning about spatio-temporal continuity. Image and Vision Computing **26** (2008) 67–81 http://www.comp.leeds.ac.uk/qsr/pub/Bennett08imavis.pdf.
12. Goodman, N.: The Structure of Appearance. Bobbs-Merill (second edition, 1966) (1951)
13. Whitehead, A.N.: Process and Reality: corrected edition. The Free Press, Macmillan Pub. Co., New York (1978) edited by D.R. Griffin and D.W. Sherburne.
14. Halpern, J.Y.: An analysis of first-order logics of probability. Artificial Intelligence **46** (1990) 311–350
15. Bacchus, F., Halpern, J.Y., Levesque, H.J.: Reasoning about noisy sensors and effectors in the situation calculus. Artificial Intelligence **111** (1999) 171–208