Hypothesis selection for scene interpretation using grammatical models of scene evolution

R. Young, J. Kittler and J. Matas

Centre for Vision, Speech and Signal Processing,

School of Electronic Engineering, Information Technology and Mathematics,

University of Surrey, Guildford GU2 5XH, United Kingdom.

e-mail: R.Young@surrey.ac.uk

Abstract

A major bottleneck in dynamic scene interpretation is the search that is required through a database to find a model that best matches the observed data. We show that the problem can be alleviated if the object model selection is controlled by a scene evolution model. We adopt a grammatical model to characterise objects and events in a dynamic scene which can be used to generate visual expectations within a particular context. The object hypotheses can be accepted without further search of the database provided a measure of the goodness of fit of the match between the selected model and the visual data falls below a threshold. In this paper we present experiments for determining the necessary thresholds for the model hypotheses testing using the recognition method described in [8], as well as for assessing the subsequent performance of the scene interpretation system with and without the constraining grammar.

1. Introduction

The visual world and associated dynamic events are, in many scenarios, highly regular. Prior knowledge of that regularity can reveal expectations which can be used to control and improve visual data processing. The experiments presented in this paper show how the speed of processing of scene interpretation can be enhanced by constraints imposed by a grammatical model of scene evolution. The grammar consist of facts which define the knowledge about a particular domain and rules which define the transitions between states within that context. The idea of using grammatical models of scene evolution was presented earlier [1, 5]. In this paper we report on experiments carried out to confirm the conjectures concerning efficiency gains the approach affords.

Normally, search is required through an entire model database to verify an object hypothesis. However, this can be avoided if the likely model hypotheses can be predicted. The acceptance of selected hypotheses requires the knowledge of model verification thresholds. Experiments to determine such thresholds are first presented. We then evaluate the advocated approach on a series of experiments involving real time vision processing.

2. Scene Evolution and its grammatical model

The spatial structure of the world has been the main focus of computer vision research for decades. The world is also organised temporally, in that the evolutionary structure of scene events exhibits substantial regularity. Particular constraints on the temporal structure can be predicted if the context of the scene is known and understood. The concept of *breakfast* immediately conjures up not only a particular set of objects, such as cup, saucer, sugar bowl, milk jug, teapot and cereal box, but also particular types of events, such as placement of a cup on a saucer. We shall consider a breakfast table scenario as an experimental setting to convey the benefits of exploiting temporal context in visual processing. Other research has focused on scene interpretation with spatio-temporal [7] and motion models [2, 3, 6], but mainly for the purposes of event description not performance enhancement.

The state of a scene can be either **static** or **dynamic**. In a dynamic scene objects are in motion and represent the events that take place in a transition between two static states. Events can fall into a number of different classes. A transition from a dynamic state to a static state represents a *placement* (or removal) of an object. In our breakfast scenario pouring and stirring are *motion* events. Another type of event that is of importance is a *geometric* event, such as the vertical alignment between a teapot and a cup, representing additional evidence for the hypothesis of tea-pouring.

The regularities of dynamic events and objects associated with a particular scene context can be reflected in a set of rules and facts which collectively form a grammar [5] of probable scene evolution. Here's an example of the rules (in Backus-Naur form),

```
<SET_CUPX> := <SET_SAUCER>, <SET_CUP>
<SET_SAUCER> := enter_fov(saucer), place(saucer)
<SET_CUP> := enter_fov(cup),
             alignment(cup, saucer), place(cup)
```

where the lower case items represent terminal symbols that are facts detected by the low-level visual processing modules. Each rule represents a sequence of likely events. Once one event is detected, processing resources can be concentrated on the looking out for the next event in the list. In this way wasted processing can be avoided by ignoring, or de-prioritising, hypotheses related to less likely events.

3. Object recognition

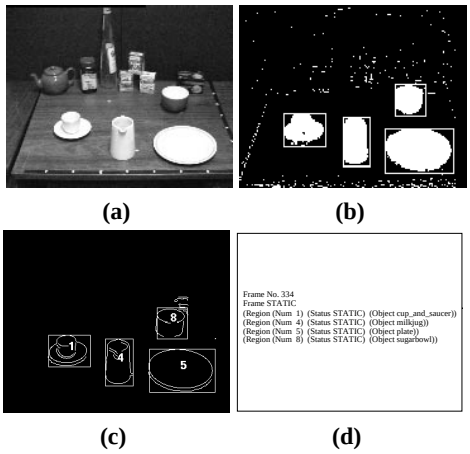

Figure 1. Processing chain a) One image in sequence b) Colour difference regions c) Edges within regions of interest d) Symbolic interpretation of regions

The object recognition approach used in the present system has been reported in detail elsewhere [8]. It uses a dedicated recognition engine for each type of object that can be found in a breakfast scenario. In particular we can cope with plates, saucers, sugar bowls, cups and milk jugs. The recognition scheme assumes some prior knowledge and constraints. All objects must be placed on a common, flat ground plane. The transformation between the camera coordinate system and the ground plane coordinate system must be known (established through calibration). The recognition procedure adheres to the processing chain shown in figure 1.

Regions of interest are determined by comparison of the current image with a background image of a static tabletop scene. Any areas which show a significant chromatic difference [4] are likely to represent new objects or events and are, therefore, deemed interesting. The outline of objects within the regions of interest are extracted and compared with the projection of three-dimensional models onto the image plane. The model that returns the closest goodness of fit value is taken as the identity of the current object. A grammar of facts and state transition rules hypothesises and prioritises expectations of probable future events from the database of current objects in the scene.

The standard procedure for finding which model is appropriate to the observations is to invoke all the relevant hypotheses based on salient features extracted from the visual data by low-level vision computing. However, such an approach is often incapable of discriminating between models in the database adequately and consequently a large number of hypotheses have to be evaluated. The aim of our approach is to narrow down the large range of possibilities by exploiting our prior knowledge of scene evolution and in this way reduce the complexity of processing.

4. Experimental Setup

In order to demonstrate the computational benefit of an approach we have performed a number of scene interpretation experiments where the observed scene evolution adhered to a grammatical model to a varying degree. The experiments were performed with a real-time vision system, on-line with three different grammars.

In order to maintain a constant and stable viewing position a JVC TK1070E camera was mounted on a PUMA762 robot arm and trained on the tabletop scene. Processing was performed on a Silicon Graphics server with live images input from a Sirius image grabber. Although the grabbing and processing rate of, 1 - 5 frames per second is below frame rate it is still fast enough for the experiments to be carried out in real-time.

Two sets of experiments were carried out. To facilitate a fully automatic operation it was first necessary to determine suitable hypothesis testing thresholds for each object model. This was accomplished by placing each of the five test objects in different positions in the field of view covering the tabletop scene. At each point the values of the match between the observed object and every model in the database were measured and recorded.

The grammar, of rules written in CLIPS, defined the *order* in which objects were expected to arrive in the scene. The second set of experiments involved a sequence (the same each time) of objects being placed in the scene. Three different grammars were used to generate the expected objects. One defined the same ordering as the sequence, one a

minor change to the ordering and the third the complete opposite sequence. The sequences were also run without using any grammar. The results recorded in each case were the average number of model comparisons made and the average CPU time, for a successful database search and recognition.

5. Results

The results of the experiments are shown in figure 2 and table 1. Figure 2 shows the graph, for four of the five objects, of the number of positions in which the goodness of fit falls within a particular interval, for each of the five models. The solid line represents the distribution of the match values for the correct model, whereas the dotted line is the mixture histogram of match values with all the other models. From these results, match thresholds for the saucer, cup_and_saucer, milk jug, plate and sugar bowl were derived as 4.10, 2.35, 3.50, 4.60 and 4.00, respectively.

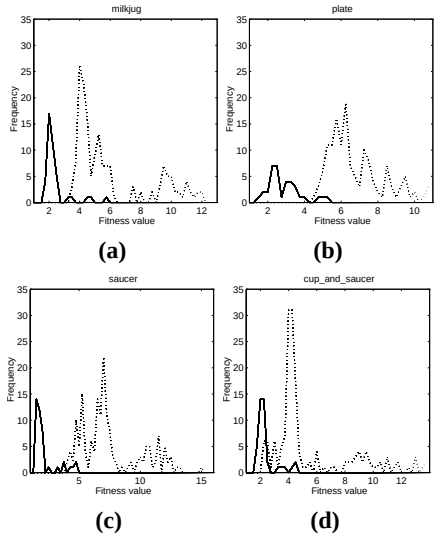

Figure 2. Determination of match thresholds

Table 1 shows the number of model comparisons and corresponding cpu times (seconds) for successful recognition of new objects placed in the scene. The rows show the results for situations where the deviation of the grammar defined sequence from the actual sequence was nil, minor deviation and major deviation, as well as with no predictive grammar.

When the predicted and actual sequences are the same only one model need be checked with the corresponding gain in processing time. As the actual and predicted sequences deviate more the performance deteriorates. The process time for the major deviation is more than twice as slow as no deviation. This time is, of course, for a database of five objects and would therefore increase as the size of the database grows whereas the time for nil deviations would remain constant. Although the situation did not arise in our experiments, there may be cases where objects are misidentified due to the overlap in goodness of fit values (see figure 2 (d)). Recovery from such errors will be a goal of future research in this area.

Model comparisons and CPU times		
Deviation	Models	CPU
None	1.0	0.07
Minor	1.75	0.09
Major	3.7	0.17
No grammar	2.8	0.13

Table 1. Results of grammar experiments

6. Conclusions

We have demonstrated that substantial gains in the the speed of scene interpretation can be achieved by means of hypothesis generation based on prior temporal contextual knowledge of scene evolution, encapsulated by a grammatical model. These results confirm the conjectures made in an earlier model [1, 5]. They show what an important role temporal context can play in visual processing, especially in the case of dynamic scenes with strong temporal ordering of events.

References

[1] H. Christensen, J. Matas, and J. Kittler. Using grammars for scene interpretation. In *ICIP*, pages 793–796, 1992.

[2] D. Koller, K. Daniilidis, T. Thorhallson, and H.-H. Nagel. Model-based object tracking in traffic scenes. In *ECCV*, pages 437–452, 1992.

[3] H. Kollnig, H.-H. Nagel, and M. Otte. Association of motion verbs with vehicle movements extracted from dense optical flow fields. In *ECCV*, pages 338–347, 1994.

[4] J. Matas. *Colour-based Object Recognition*. PhD thesis, University of Surrey, Guildford, Surrey GU2 5XH, 1995.

[5] J. Matas, J. V. Kittler, J. Illingworth, L. Nguyen, and H. I. Christensen. Constraining visual expectations using a grammar of scene events. In I. Plander, editor, *International Conference on Artificial Intelligence and Information-Control Systems of Robots Singapore: World Scientific*, pages 81–92, 1994.

[6] M. Mohnhaupt and B. Neumann. Understanding object motion: Recognition, learning and spatiotemporal reasoning. In *Towards Learning Robots*, pages 65–91. MIT, 1993.

[7] A. Toal and H. Buxton. Spatio-temporal reasoning within a traffic surveillance system. In *ECCV*, pages 884–892, 1992.

[8] D. Yang, J. V. Kittler, and J. Matas. Recognition of cylindrical objects using occluding boundaries obtained from colour based segmentation. In E. Hancock, editor, *British Machine Vision Conference*, pages 439–448, 1994.