Sri Lanka Association for Artificial Intelligence (SLAAI)
Proceedings of second Annual Session
19th February 2005 – Colombo

Towards building a Cognitive Vision System for learning in Behavioural Models from Symbolic Data using Qualitative Spatio-Temporal Relations

D. D. M. Ranasinghe,
Department of Mathematics & Computer Science, The Open University of Sri Lanka, Nawala, Nugegoda, Sri Lanka
menaka_dul@yahoo.com

A.G.Cohn,
School of Computing, University of Leeds, LS2 9JT, UK,
agc@comp.leeds.ac.uk

A. S. Karunananda
Department of Information Systems and Computing, Brunel University, Uxbridge, Middlesex, UB8, 3PH, UK
asoka.karunananda@brunel.ac.uk

Abstract

Research has been carried out to develop a cognitive vision computer system that will be capable of perceiving, reasoning and learning through visual inputs. Our approach is based on the assumption that robust qualitative spatio-temporal relations extracted from visual data can be used to successfully implement cognitive vision systems that behave like humans with reasoning and learning abilities. In our ongoing research, placing covers on a dinner table has been analysed and have been identified some basic robust qualitative spatio-temporal relations such as rightof, leftof, frontof, etc. Prolog has been used for implementation of analysis of spatio-temporal relations, while extraction of relevant rules from relations has been implemented with the help from Progol, which is a many sorted language for inductive logic-programming that implements learning by examples. The final goal of this project is to make a computer-based cognitive vision system capable of learning from more comprehensive set of examples and carry out complex reasoning on spatio-temporal visual data to learn behavioural models.

1. Introduction

One of the fundamental abilities of human beings is to recognize, learn and conceptualise knowledge from visual inputs. We categorise the objects that we happen to see in the environment into semantic groups and extract relevant knowledge about them. These semantic groups can be generated according to our intention of what knowledge that we want to build from visual inputs. Another unique feature of humans is, given the same visual input the generated knowledge at different time points is different. This is because humans interpret what they see based on the prior diverse knowledge and experiences they have about the world; hence this generates the continuous process of incremental and adaptive learning. In addition, humans have the capability of applying the robust concepts/rules that they have already acquired to learn and adapt to new situations. This whole process of cognition can be interpreted as generation of knowledge on the basis of perception, reasoning, learning and prior models of the things.

In this research we intend to adapt the fundamental ability of humans explained above to learn from visual inputs, in the process of building computer systems to behave/reason like humans. We argue that in an unfamiliar environment, humans tend to abstract qualitative relations among the constituents of the environment, as these seem to be more robust. This is a key feature that can be manipulated to build autonomous cognitive vision systems and at present this feature is not exploited to a great extent [9,12,13]. Therefore, in our research we exploit qualitative spatio-temporal relations in extracting robust concepts/rules. These learned spatio-temporal

concepts/rules can be used to learn the protocol behaviour; hence behavioural models for unknown environments can be predicted.

The rest of this paper is organised as follows. Section 2 carries a brief description on Cognitive Vision Systems with the research questions that is being analysed. Section 3 presents research design. Section 4 reports on so far results and findings. Section 5 is a conclusion and presents further work.

2. Cognitive Vision Systems

We humans tend to learn many things from observations: it can be by observing another human performing the same task or a simulated agent doing the same thing. Consider the scenario of a more mature child building a house using blocks. A younger child observing this dynamic scene will tend to mimic the same thing. In doing so, he tries to learn how to place the appropriate blocks at correct places. We argue that this is learning of spatial relations between objects and this learning is more qualitative than quantitative. In addition, the child will learn before placing the roof he has to complete the walls of the house; hence the temporal knowledge is abstracted.

Based on this scenario we can adapt many things in the process of building fully autonomous systems, such as learning from observations, abstraction of qualitative spatio-temporal relations, adaptive/incremental learning according to intentions. The Cognitive Computer Vision research area exploits this idea of learning from visual inputs.

Cognitive Computer Vision is regarded as the area that deals with vision data with respect to the cognitive process of knowing, understanding and learning about the things that we happen to see [14]. These systems include facilities for acquiring information from the outside external world through learning or association and produce a response to appropriate percepts. Therefore these facilities should incorporate understanding, knowing, and learning from the data of the outside world that is being captured and build semantic models of the environments. In order to embed these capabilities, the development of cognitive vision systems proposed in the CogVis (Cognitive Vision systems, IST-2000-29375) project is divided into four work packages namely, i) recognition and categorization of objects, structures and events, ii) Reasoning and interpretation about scenes and events, iii) learning and adaptation, and iv) control and integration [14].

In work package one, it is intended to build computational systems, which are capable of recognizing and categorizing objects and events in a natural environment. Previously this task was mainly considered as recognition based on prior defined models. At present much research is done towards building systems that can recognize and categorize objects and events in unrestricted environments with spatial and temporal extensions. The objective of work package two is to develop conceptual structures for high-level knowledge and reasoning processes for scene understanding. To develop life like vision systems they should be capable of going beyond mere object recognition. Since objects are only a part of a whole scene the cognitive vision systems should be capable of integrating the identified parts to a whole picture to act as an agent carrying out a particular task. Work package three addresses the issue of how knowledge about objects and events are acquired and maintained. A cognitive vision system fully interactive in a particular environment should be able to learn from its experiences and adapt to situations as humans do. Therefore it is considered that the learning should be continuous by dividing it into appropriate phases in an unsupervised or exploratory manner. In work package four distributed methods for attention and control with mechanisms for system integration are considered. Since it is considered that cognitive vision systems should be able to operate in a continuous manner it should have capabilities to balance visual attention and control of its interaction in integration with such other systems. In this paper we do not expect to do any detailed study about the work involved in these packages. Further information can be obtained from [14]. With the integration of the work in each of these packages it is intended to develop cognitive vision systems in realistic settings. Further, it is considered that cognitive vision systems should be able to incorporate vision, which is not instantaneous and generate knowledge by generating the models of the perceived world under finite resources without being constrained by closed world assumption. In our research we mainly intend to work on work package two.

Most of the early applications of cognitive vision systems were mainly focused to generate verbal descriptions about the observed scene. They mainly used an in built set of vocabulary with an integrated knowledge base hence the systems were task dependent as well as

situation dependent. Since this was a great drawback in the process of building autonomous agents that can be implemented in realistic settings the need arise to go beyond mere interpretation of scenes.

Vision is a process that evolves over time, gathering information to build up knowledge; hence it operates in a spatio-temporal context. From the visual information gathered, knowledge can be developed about the environment in terms of its geometry, by generating semantic labels for events and entities in the environment. It is argued that when the environment is understood, it is possible to generate explicit descriptions in terms of objects, structures, events and their relations. These learned concepts provide the basis for developing an explicit behavioural model in a particular environment either by generating an action or a form of communication that enables autonomous behaviour.

Motivated by the previously explained scenario, the research questions that we will be analysing in this research are:

- How can cognitive systems capture and manipulate information through their percepts?
- How can these systems learn from examples?
- Is it possible for a cognitive vision system to represent its environment predominantly in qualitative terms?

We argue that the usability of task specific agents is restricted to the environment they have been designed for; therefore such a development is less effective. Therefore we intend to pay more attention on learning of underlying robust relations of the present environment. By doing so it may be possible to develop systems that can exhibit protocol behaviour even in an unknown environment.

3. Research Design/Material and Methods

To have a cognitive system that can respond to percepts and which is fully immersed in the real world it should be able to learn appropriately from observations of its environment.

In the process of developing such a cognitive system, the dynamic scene is captured by a video image capturing process in a frame-by-frame basis [7]. Given an image sequence with a moving blob, the difference image is obtained from the previous and following frames. The arithmetic mid points of the bounding boxes of identified components is then used as seed mid points for the region-growing algorithm. An advantage of this approach is that it does not require knowledge about the background, but assumes uniformly coloured objects. Perceptual groups are identified using an attentional mechanism subsequently generating a set of symbolic data. The generated symbolic data contain the frame number, type of the object, object ID, centre coordinates, the bounding box [1], similarity measure and information regarding whether the object is moving or not. The object ID is particularly useful for reasoning about objects whose appearance changes during a scene. For each object similarity measure is calculated with respect to reference objects.

3.1 Learning qualitative spatial and spatio- temporal relations

Qualitative reasoning can be regarded as one form of capturing everyday commonsense knowledge. This has many advantages over quantitative reasoning due to the ability of making inferences with fewer calculations even in situations with incomplete knowledge. Further, spatial reasoning in our everyday interactions with the real physical world is mainly driven by qualitative abstractions rather than complete priori quantitative knowledge. Therefore we primarily intend to abstract qualitative spatio-temporal knowledge of the constituent objects of the captured dynamic scene.

For a prototype, we are considering the scenario of laying covers in a dinner table setting. This dynamic scene, which is represented by a set of symbolic data, is analysed in a top down manner considering intentions. We adopt the techniques used in qualitative spatial reasoning to exploit the spatial concepts in symbolic data [2,4,6].

Some of the captured video films contain multiple covers. Therefore the first step in the analysis is to cluster the objects that belong to one cover. This is done by the use of qualitative distance relations combined with orientation [15]. A vector that connects the primary object and the reference object gives the spatial positional information of a primary object. This representation yields a local orthogonal grid

[1] It is assumed that the objects are identified separately on a frame-by-frame basis; for exceptions refer to [1].

with 15 qualitative relations that form a conceptual neighbourhood.

Initially we represent the extended objects with centre points because they do not change their form during the movement and the objects are spatially disjoint. Using the centre points entire movement trajectory of a single object is analysed. The bounding box information is used later on to do the same analysis.

These spatial relations are further analysed to incorporate temporal variations. An Inductive Logic Programming ILP [10] tool called Progol [11] is used to draw out the underlying rules/concepts of the learned qualitative relations. A meta Prolog program executes in between, transforming the spatial relations into an appropriate format for Progol.

3.2 Learning behavioural rules using Progol

To learn behavioural rules we use Progol, an Inductive Logic Programming (ILP) tool. ILP is a machine learning technique based on a first order knowledge representation mechanism that implements methods for inductive generalization as a logic programming system. Given background knowledge and a set of observations ILP aims to induce a logic program which generalizes the observations. Inductive logic programs learn the general concepts or hypotheses in the form of rules, which are referred to here as behavioural rules, that best explain the given examples. In this domain we only have positive examples. The ILP system we use has a setting for learning from positive only data. For a more detailed overview of ILP refer to [10].

Progol is an ILP tool that generates logic programs in the form of hypotheses/rules in the light of given examples and background knowledge. In brief, Progol works as follows. For each positive example the most specific Horn clause is generated according to the user declared mode declarations. These mode declarations impose restrictions on generalisations. The initial most specific clause is further being checked against each remaining positive example for improving the generalisation power to subsume the highest number of positive examples. Early applications of Progol mainly applied for finding new facts about molecular biology.

To understand this in simple terms next we will consider an example from [11]. Say we want Progol to construct a definition for the relation 'aunt_of' then examples should be given for the relation such as:

```
aunt_of (jane,henry).
aunt_of(sally,jim).
```

The background knowledge can be about relations such as 'parent_of', 'father_of', 'mother_of', etc.

```
Parent_of(Parent,Child):-
        father_of(Parent,Child).
..
father_of(sam,henry).
..
```

and so on.

In addition a list of types and mode declarations have to be given. Types describe the categories of objects (numbers, names, lists, etc.) about the considered scenario. In this example we need only the type *person*, since all objects given in the relations are of this type.

```
person(jane).
..
```

The modes describe the relations (predicates) between objects; types of the objects and the form of these atoms can be in the hypothesised clause. More specifically *modeh* statements indicate the predicates that come in the head of the generalized rule while *modeb* statements indicate the predicates that come in the body of the generalized rules. Such as:

```
modeh(1,aunt_of(+person,+person))?
modeb(*,parent_of(+person,-person))?
```

According to the *modeh* statement the head atom has predicate symbol 'aunt_of' and has two variables of type *person*. The '+` sign indicates that an argument is an input variable while '–` indicates an output variable, and a '#` indicates that a constant should be placed at the particular position in the hypothesis. In mode declarations the numerical value or the '*' is known as the *recall* value. This bounds the number of alternative solutions. Here in *modeh* declaration it is 1 because for a given predicate of 'aunt-of' has a unique answer either 'yes' or 'no'. With all these information Progol will construct a generalized rule that will cover maximum number of examples such as:

aunt_of(A,B):-parent_of(C,B), sister_of (A,C).

Further, Progol is a system that can learn from positive only examples and this is particularly important in our table setting scenario because negative examples are not a natural occurrence

that match with the underlying intentions. Even in real life also this characteristic suits well because we usually come across understanding visual scenes without active guidance to the learning process. By giving the background knowledge the hypothesis searching mechanism is guided rather than merely searching space of syntactically legal hypothesis; hence efficiency of the system is enhanced.

Moreover Progol needs lot of redundancy which can be seen as an advantage because we believe that even humans tend to learn things better after being exposed to the same experience several times. The requirement to manually give mode declarations in Progol is tedious and makes the learning less autonomous. Therefore, here we intend to guide the mode declaration process with the use of a descriptive inductive logic-programming tool HR by learning integrity constraints [5]. The learned robust spatial temporal rules/concepts are the protocol behavioural rules that can be converted to actions. A Prolog meta program relating to corresponding actions can interpret these learnt rules.

4. Results/Findings

The dynamic scene of placing covers on a dinner table is considered for developing a proto-type of the proposed system. Currently some of the basic qualitative spatio-temporal relations such as leftof, rightof, and frontof are analysed using Prolog for a particular time point considering centre values as well as bounding box values. The whole area of a single cover is divided into six tiles based on cardinal direction model [8]. A dinner plate is considered as a reference object and it is assumed to lie in the center tile. A modification is adapted for [8] by omitting three tiles that lie below the plate with the assumption that the dinner plate lies in a very closer position to the person who is being served and there is no room for another set of tiles below than the dinner plate. The same relations are analysed to incorporate the variations with the time. Using Progol, relevant rules are extracted for these relations. One such rule is:

Frontof(A,B):- center(A,[C,D]), center(B, [E,F]), D=< F.

Where A, B are objects and. C, D and E, F are the corresponding centre coordinates of the two objects. This rule correctly identifies the conditions that should be satisfied object A to be in front of object B (according to the data set value of y axis decreases when moving forward). The results obtained from centre analysis as well as bounding box analysis will be compared against a hand coded set of perfect rules to identify which analysis yield more robust results.

5. Conclusion & Further Work

We have presented our research work on the development of a cognitive vision system that exploit qualitative robust spatio-temporal relations to learn inductively from visual inputs thereby leading to reasoning in complex visual systems. The research is still at its early stages and more work is needed to enable the system to learn from a more comprehensive set of visual inputs pertaining to real world scenario.

The initial idea of this work is originated based on the work done at the University of Leeds in the CogVis project (Cognitive Vision systems, IST-2000-29375). So far, work at University of Leeds has made little use of qualitative spatio temporal relations [9,12,13]. In our work we propose to exploit qualitative spatio-temporal relations more, e.g. by abstracting conceptual neighbourhood relations in RCC5 [3] model. Since our work is still at a very early stage we need further investigation how incremental/adaptive development of these qualitative relations can be utilised in predicting the protocol behaviour in an unknown environment.

6. References

[1] Bennett B., Magee D., Cohn A.G., Hogg D., Using Spatio-Temporal Continuity Constraints to Enhance Visual Tracking of Moving Objects, Proc. 16th European Conference on Artificial Intelligence (ECAI-04), 2004.

[2] Cohn A., G., The Challenge of Qualitative Spatial Reasoning, ACM Computing Surveys, Volume 27 (3), 1995, pp 1-30.

[3] Cohn A.G., Bennett B., Goday J., Goots N. M., Qualitative Spatial Representation and Reasoning with the Region Connection Calculus, Geoinformatica, 1, Kluwer academic publishers, pp 1-44, 1997.

[4] Cohn A. G., Hazarika S. M., Qualitative Spatial Representation and Reasoning: An Overview, Fundamenta Informatica 46(1-2), 2001,pp 2-32 IOS Press.

[5] Colton S., Muggleton S., ILP for Mathematical Discovery, In Proceedings of the 13th International Conference on Inductive Logic Programming, 2003.

[6] Hernandez D., Qualitative Representation of Spatial Knowledge, Number 804, In Lecture Notes in Artificial Intelligence, Springer-Verlag, 1994.

[7] Hongeng S., Unsupervised Learning of Multi Object Event Classes, Technical Report FBI-HH-B-257/04, Informatik, University of Hamburg, 2004.

[8] Kor A. L., Bennett B., Composition for Cardinal Directions by Decomposing Horizontal and Vertical Constraints, workshop in Foundations and Applications of Spatio-Temporal Reasoning, AAAI Spring Symposium, 2003, pp 39-45.

[9] Magee D., Needham C., Santos P., Cohn A., Hogg D., Autonomous Learning for a Cognitive Agent using Continuous Models and Inductive Logic Programming from Audio-Visual Inputs, Proceedings of the AAAI workshop on Anchoring Symbols to Sensor Data, 2004.

[10] Muggleton S., Raedt L., Inductive Logic Programming: Theory and Methods, Journal of Logic Programming, 19, 1994, pp 629-679.

[11] Muggleton S., Firth J., Cprogol4.4; A Tutorial Introduction, In Relational Data Mining, editors, Dzeroski S., Lavarac N., Springer Verlag, 2001, pp160-188.

[12] Santos P., Magee D., Cohn A., G., Looking for Logic in Vision, Proc. Eleventh Workshop on Automated Reasoning, 2004, pp 61-62.

[13] Santos P., Magee D., Cohn A., Hogg D., Combining Multiple Answers for Learning Mathematical Structures from Visual Observation, Proc. 16th European Conference on Artificial Intelligence (ECAI-04), 2004.

[14] Technical Annex 1: Cognitive Vision Systems, Description of Work, IST-2000-29375, 2001.

[15] Zimmermann K., Freska C., Qualitative Spatial reasoning Using Orientation, Distance, and Path Knowledge, Applied Intelligence 6, 1996, pp 49-58.