Qualitative Realization of the Visual World

D. D. M. Ranasinghe,

Department of Mathematics & Computer Science, The Open University of Sri Lanka, Nawala, Nugegoda, Sri Lanka

E-mail: menaka_dul@yahoo.com

Abstract

Visual capability is one of the strongest senses of perception of human beings. Humans tend to represent the visually perceived world mainly in a qualitative manner and thereby attain considerably high accuracy in reasoning and prediction. Even though this is an innate ability of humans, embedding this feature in the development of cognitive vision systems has been a research challenge. A research has been carried out to develop a system that is capable of learning qualitative rules that underlies in the arrangement of the observed visual scene. A set of symbolic data generated from a dynamic visual scene that comprise object movement is considered as the input to the system and by analyzing object-object qualitative spatial and temporal representation and reasoning mechanisms the system generates the underlying set of rules of the scene. As an application the system produces sketch images of the observed scene. This work has great potential in developing agents that can be used for autonomous learning from visual scenes in a manner closer to human learning from visual scenes.

1. Introduction

Visual ability is one of the richest and most developed avenues of information gathering for humans. Therefore learning from visual scenes is a natural cognitive process and we humans, tend to represent this perceived real world mainly in a qualitative manner rather than in a quantitative manner. Even though quantitative measures yields high precision in representation and reasoning, use of phrases such as *the post office is left to the school* in describing a location, *take the second turn to right side* when describing a direction and *keep the saucer before keeping the cup on top of it* are some common examples for usage of qualitative knowledge in day-to-day life. Further, qualitative representation and reasoning is a good alternative when exact information is missing or when the quantitative manipulations are costly. In addition, the limited mental ability of an average human in performing quantitative manipulations is also another factor for using qualitative knowledge.

Hence, when observing an evolving scene we tend to abstract various qualitative features such as movement, size, brightness, position, and orientation etc of the objects in the scene. It is evident that most of these features are either spatial concepts or temporal concepts. In general much of the knowledge about space and time is qualitative than quantitative [1].

By observing an evolving scene humans tend to develop a conceptual understanding about what is taking place and this understanding is conditioned by exposing to similar situations. Depending on the intention of the observer this abstracted knowledge can be organized to derive new knowledge or update the existing knowledge. In doing so one has to learn the underlying rules of the scene.

Based on the above philosophy we have developed a computer-emulated system that exploit qualitative spatial and temporal representation and reasoning mechanisms on a set of symbolic data obtained from a visual scene and thereby generate rules of the scene using Inductive Logic Programming (ILP).

The rest of the paper is organized as follows. Section 2 carries an overview of cognitive vision systems that exploit qualitative representation and reasoning mechanisms. Section 3 is on the theoretical foundations adopted and section 4 reports the approach taken. Section 5 is on design and implementation while section 6 carries a discussion about the results. Finally section 7 is on conclusion and further work.

2. Qualitative representation and reasoning of visual scenes

The area of cognitive computer vision systems is devoted for learning from cognitively enabled vision and is an active research area in the field of AI [2]. Cognitive computer vision systems deal with vision data with respect to the cognitive process of knowing, understanding, and learning about the things that we happen to see. Hence it has facilities for acquiring data from the outside world through learning or

association and produces a response to appropriate percept. Cognitive vision computer systems that can be fully embedded in the environment need to have facilities for automatically acquiring data, process them and develop conceptual models of the environment in such a manner as humans do.

Badler's work is considered as one of the earliest attempts to learn qualitative models of visual scenes [3]. He conceptualised the observed scene with a hierarchy of motion concepts such as behind, after and so on. Due to the limitations in technology, obtaining these concepts automatically was not a feasible task at that time. Hence these concepts were given beforehand and this reduced the flexibility of the system.

In the views project spatial representations were considered as cells with topological properties, which supports the topological reasoning required by the system [4]. Due to the use of quantitative reasoning mechanisms with a coordinate system lead to hand generation of spatial regions. This has limited the flexibility of the system. Development of spatial regions was automated by using qualitative spatial and spatio-temporal (s-t) reasoning methods in [5]. Here the use of qualitative spatial and s-t regions were limited to conceptual clustering of low level input data and has not gone to the extent of learning any rules to build models of the observed scene.

Description logic was used to learn scene semantics by analysing qualitative spatial and s-t relations in [6]. Due to the pre determination of the learnt relations the work cannot be generalized to learn in another situation. By exploiting spatial and s-t relations context specific rules are learnt in [7],[8]. Even though they have adopted a similar approach as ours the application of spatial and s-t relations was limited to identification of perceptual groups and the notion of presence and absence of objects. Further any qualitative object-object spatial or temporal relations were not analysed. We argue that cognitive systems that can be fully embedded in the environment should posses capabilities in representation and reasoning in such a manner as humans do.

3. Theoretical Foundations

Since spatial reasoning in our everyday interactions with the physical world is generally driven by a qualitative abstraction rather than using exact quantitative measures, there are various theories developed to address qualitative spatial and s-t representation and reasoning [9]. Since we account for the spatial extent of the objects we do not consider any point-set topological theories here. Hence Region Connection Calculus (RCC-8) theory [10] is exploited to determine the type of connection between any two regions and Allen's interval algebra [11] to account for temporal variations in object movement.

A Region Connection calculus (RCC-8)

The theory describes the primary relation of any two spatial entities in the form of connection. Therefore the basic dyadic relation is C(x,y), which reads as x connects with y. Based on the type of connection a set of eight jointly exhaustive and pair wise disjoint (JEPD) relations are derived as shown in fig.1.

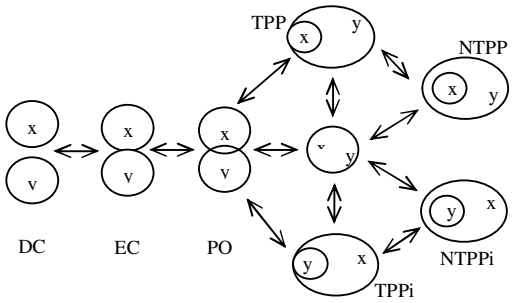

Fig. 1. Relations of RCC-8 calculus and their continuous transitions

The abbreviation DC (x, y) is read as x is disconnected from y. Similarly the other abbreviations are EC-externally connected, PO-partially overlaps, EQ-equal to, TPP-tangential proper part, TPPi-inverse of tangential proper part, NTPP-non tangential proper part and NTTPi- inverse of non tangential proper part.

Further this theory illustrates that when any two spatial entities move they can move sequentially only in adjacent locations. Our adaptation of RCC-8 theory to describe the type of connection between any two moving objects is described in section 5 *c*. To account for movement we employ Allen's interval algebra and an outline of the theory is given in the next section.

B Allen's Interval Calculus

Allen's interval calculus describes the temporal relation between any two moving objects. According to Allen's notion there are 13 JEPD relations, and for any two intervals exactly one of the relations holds as shown in fig.2.

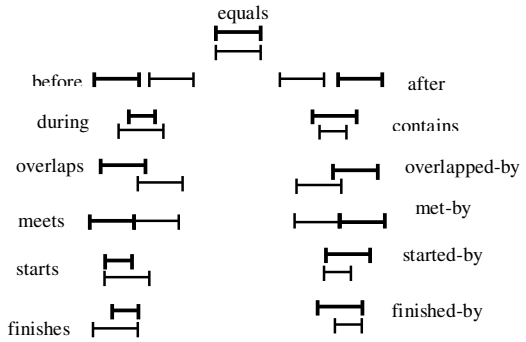

Fig. 2. Temporal Intervals in Allen's calculus

The Allen's relations are defined considering the end points of the intervals and allow movement only in adjacent time stamps. Thus this helps to anticipate the future as well as describing the relations of a third interval with the use of transitivity properties. In next section we will look at the approach taken to implement the proposed framework.

4. Approach

The real world scene is captured by a video capturing process in a frame-by-frame manner and converted to symbolic data using an attention mechanism [12]. We assume this is done before hand.

Initially the symbolic data set is clustered into perceptual groups based on a reference object. An object with salient features is often regarded as a reference object [1]. Some of the salient features are size, brightness, movement, etc.

Since we account only for indoor scenes of object movement the representing world can be regarded as a small-scale space. In small-scale spaces topological and orientation relations provides a restricted form of positional information, which describe the arrangement of the objects [1]. Orientation relations describe where objects are kept relative to one another and can be defined with three basic concepts, the primary object, the reference object and the frame of reference [13]. An intrinsic frame of reference is assumed by considering the characteristic direction of movement of the reference object. The object in which the position has to be determined is called the primary object.

The earlier described RCC-8 theory is used to identify the topological relations of the boundaries of the objects [1]. The theory is adopted in such a way to determine whether objects are apart from each other, any two objects touch each other or one object is on top of another object. Even though TPP, NTPP and their inverses are realistic situations in object movement we treat all of them as EQ. Because in laymen terms, those situations can be considered as situations, where, one object is on top of another object.

In situations where an object A is kept on top of object B a human knows intuitively that object B has to be kept prior to object A because vies versa is an unrealistic situation. The only possible way of learning such concepts for machines is by considering the time factor. Similarly we too account for the order of object placing by considering the Allen's interval calculus for time.

5. Design and Implementation

A dynamic scene of setting covers in a dinner table is used as a prototype model to implement the proposed framework. Fig. 3 is a snap shot of a one such captured scene.

Fig. 3. A Snap shot of the table setting scenario

The symbolic data set contain information about frame number, type of the object, object ID, center coordinates, the bounding box, similarity measure and information regarding whether the object is moving or not.
The design of the system is given in fig. 4. There are two main modules in the system namely Qualitative Knowledge (QK) module and the ILP module.

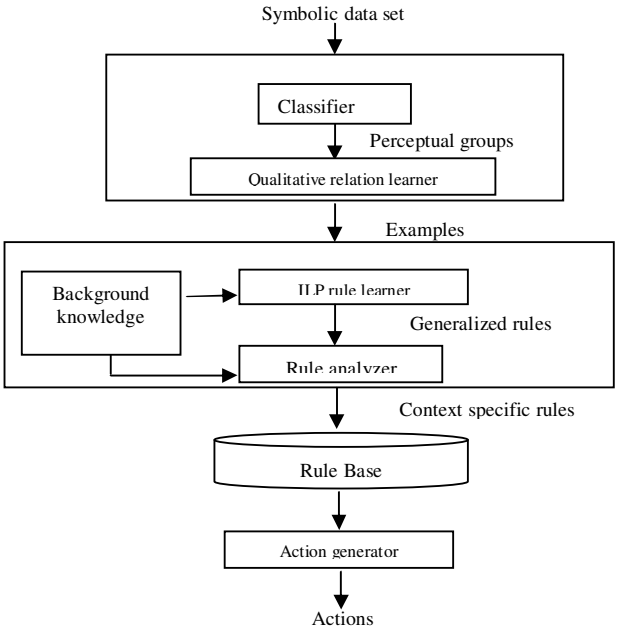

Fig. 4. Overall system design

The two modules are further sub divided into smaller modules with a specific task in hand and are explained below.

5.1. Identifying perceptual groups

To account for the dynamic nature of the scenes we assume non-monotonic reasoning methodology by assuming that new evidence can invalidate the past conclusions [14]. Accordingly the first appearing object is considered as the reference object and when a second object comes, the previously assumed reference object is not used any more. Then the larger object in size is considered as the reference object according to our adaptation. Therefore the reference object is calculated dynamically for each and every frame. The other objects that appear in the frame are clustered, based on a distance constraint of three times the bounding box of the reference object as the limiting distance for a single cluster [1]. The clustering that appears when the scene becomes static is considered as the final grouping of perceptual groups.

5.2. Relative orientation of objects

According to the prototype the largest object, *plate* is considered as the reference object. Based on the reference object eight distinct relations are used: *front* (F), *back* (B), *left* (L), *right*(R), *le*ft-back (LB), right-back (RB), left-front (LF) and right-front (RF) to determine the relative orientation of a primary object. In doing so, we employ a cardinal direction type of a grid system for the layout of the scenario [15]. The reference object is considered to be in the center tile and other primary objects that are in a single cluster fall in surrounding tiles as shown in fig. 5.

The bounding box of the tile that contains the reference object determines the size of the grid. X_{min} and X_{max} are the values of the X-axis of the left most corner and the right most corner of the bounding box respectively. So as Y_{min} and Y_{max}.

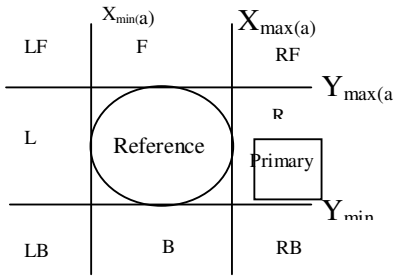

Fig. 5. Boundary line grid system

Based on this notation the orientation of a primary object (∅) can be determined by considering constraints such as:

$Front(\phi) = \{\langle x, y \rangle \mid X_{min}(r) \leq x \leq X_{max}(r) \wedge y \geq Y_{max}(r)\}$

$FrontRight(\phi) = \{\langle x, y \rangle \mid x \geq X_{max}(r) \wedge y \geq Y_{max}(r)\}$

Similarly the boundaries of other tiles too are determined.

It is noted that according to this notation sometimes one or more objects falls into the same tile. For example, according to our prototype model k*nife* and *spoon* both are on the right side of the plate. In such situations it is necessary to determine the relative orientation of the objects that fall in the same tile. Therefore, the same above explained mechanism is carried out for objects that fall with in the same tile and relative orientations are defined again. By doing so more finer orientation relations like *the knife and spoon are both right to the plate and knife is to the left of spoon* can be easily identified.

After deciding the orientation relations the next step is to identify the topological object-object relations. Topological connection between objects are identified by using RCC-8 relations.

The RCC-8 relations are identified by implementing the framework proposed in [16]. A finite set F of

cells are considered to identify the relations and the center coordinates of a primary object is checked against the bounding box values of the reference object. Relations on regions are defined as in fig 6.

DC (P,Q)	$P \cap Q = \varnothing \wedge \forall x, y(x \in P \wedge y \in Q \rightarrow \neg A(x, y))$
EC (P,Q)	$P \cap Q = \varnothing \wedge x, y(x \in P \wedge y \in Q \wedge A(x, y))$
PO (P,Q)	$P \cap Q \neq \varnothing \wedge P \square Q \wedge \square P$
EQ (P,Q)	$P = Q$
TPP (P,Q)	$P \subseteq Q \wedge \exists x, y(x \in P \wedge y \notin Q \wedge A(x, y))$
NTPP (P,Q)	$P \subseteq Q \wedge \forall x, y(x \in P \wedge A (x, y) \rightarrow y \in Q)$
TPP_i (P,Q)	$Q \subseteq P \wedge \exists x, y(x \in Q \wedge y \notin P \wedge A (x, y))$
$NTPP_i$ (P,Q)	$Q \subseteq P \wedge \forall x, y(x \in Q \wedge A (x, y) \rightarrow y \in P)$

Fig. 6. RCC relations on regions

Even though the resulting algorithm produces all eight relations we are particularly interested in DC, EC, PO and EQ. Because in real life we often come across situations where two objects are apart from each other, externally touching each other, and one object on top of the other object etc. Therefore the objects with an EQ type of connection are regarded as objects that are on top of each other. If the connection is PO then in 3 dimension, one object is being partially occluded by the other.

With the combination of topological and orientation relations the arrangement of the objects in a particular time point can be predicted. Since our prototype evolves in a time interval it is necessary to address the whole time period to determine the order of object placing.

5.3. Object-object time relations

Time factor plays an important role when there is an order restriction of placing some objects. Using Allen's calculus we account for the end points of time intervals to determine situations such as the end time interval of placing saucer has to end before the end of the time interval of placing the cup.

Time points with significant features such as change in object movement are noted. When such significant frames are encountered three frames before and after time point is taken into consideration. Even though we are mainly interested in moving objects we still account for objects that became stationary in the immediate past since the time registry of immediate past situations are more prominent.

Therefore by combining the time variations with positional information object arrangement of the visual scene is learnt. In the qualitative relation-learning module we identify only context specific relations but to reuse the learnt relations in another context we argue that learning relations only, is not sufficient thus we need to learn the rules for the learnt relations. A brief explanation about the ILP rule-learning module is given in the next section.

5.4. ILP module

The qualitative object-object spatial and temporal relations identified in the QK module are considered as input examples for the ILP module. To learn from examples ILP methodology is selected mainly because of the ability to handle symbolic data and due to the more human comprehensible nature of the output [17].

The ILP module comprises of two main components namely ILP rule learner and the rule analyser. The ILP rule leaner is implemented with the use of an off the shelf ILP package called PROGOL [18].

PROGOL generates logic programs in the form of hypotheses/rules in the light of given examples and background knowledge. In brief, rule-generating mechanism of PROGOL is 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 generated hypothesis is the one that explains the highest number of examples.

The ILP rule learner is supported by relevant background knowledge, which guides the rule learning mechanism. This enhances the efficiency of the system because the rule generating mechanism is guided rather than exploiting a syntactically possible search place. Following is one such given example for learning orientation relation for the knife with respect to the plate.

time (t54).
orientation (rightof,[plate2, knife1],t54).
box(plate2,[51, 233, 144, 327], t54).
center(knife1, [150,280],t54).
move(plate2, 1], [knife1, 1],t54).

Bounding box coordinates of the reference object plate, center coordinates of the primary object knife and information about whether the objects are moving or not are given as background knowledge. Likewise rules are found for all orientation, topological and time relations. Since these are context specific generalised rules they can be globally used in any similar situation.

The rule analyser retains the rules that explain the highest number of examples and checks for features such as most frequently occurring rules, rules with any priority factor etc, and store in the rule base for future applications.

In summary, the system works as follows. Initially the symbolic data set is clustered into perceptual groups and qualitative object-object relations are derived in the QK module then these learnt relation examples are passed to the ILP module to generate corresponding rules. The generated rules are further analysed in the rule analyser to find out most appropriate rules that represents the considered world. As an application the system is capable of producing sketch images of arrangement of a given set of objects based on the rules learnt.

The QK module is implemented in Prolog while the ILP module is implemented in PROGOL, which is an ILP tool capable of generating rules that best explain a given set of examples. A detail explanation about the ILP rule-learning module is published in [19].

6. Results and Discussion

In this research we have developed a mechanism to adopt the natural rule learning ability of humans from visual scenes into cognitive vision systems. Thus we employed the hypothesis that in learning rules of visual scenes human tend to exploit object-object robust qualitative spatial and s-t relations. Further we incorporated human ability of learning from examples because observing an evolving scene itself is an example. The proposed framework is more suitable for learning rules from indoor scenarios of object movement.
In the prototype, only the size factor is considered in determining the reference object. Even though this works fine with our prototype scenario sometimes it may be necessary to consider another feature such as brightness.
The output of the QK module is a set of qualitative orientation, connection, and time relations at a particular time point. Following is an orientation relation learnt in the QK module.

orientation (rightof, [plate2, knife1], t54)
(1)

Relation (1) describes that the orientation of knife1 with respect to plate2 is rightof at time t54. Since this relation is highly context specific it cannot be used in another similar situation to determine the orientation relation between the plate and the knife. We overcome this limitation by learning a general rule using PROGOL. Hence the learnt rule is:

orientation(rightof,[plate,knife],A):-
box(plate,[B,C,D,E],A),center(knife,[F,G],A),
move([plate,H],[knife,I],A),F>=D,F>=B,G=<E.
(2)

In (2), orientation relation and the names of the objects are the only context specific information. Therefore unlike (1), the rule (2) can be applied in any similar situation to determine the orientation of knife with respect to plate. Similarly from the ILP rule learner, generalized rules are learnt for all the other context specific qualitative spatial and s-t relations. Therefore these rules impose a set of protocol rules that can be applied in a similar situation hence this improves the flexibility and the applicability of the system.

Here we do not consider time variations for objects with DS type of connection because usually there are no order restrictions in placing discrete objects in small-scale environments. But this may not be the case in large-scale spaces if the involved objects are comparatively different in sizes, then the larger object may have to be placed before the smaller one. If the type of connection is EQ or PO then the time variation has to be considered disregard to scale of the space. Because in 3D situations these connections are depicting objects on top of each other and objects that are occluding each other.
The generated rule set is further analysed in the rule analyser to identify the most appropriate rules and passed on to a rule base for storage for future applications. For evaluation the generated rule set is tested for soundness and completeness against a hand coded example set to test the ability of learning in the presence of noise. It is noted that there is an accuracy of about 70% of generating rules under noise.

7. Conclusion and Further work

We have developed a system that captures the natural rule learning ability of humans from visual scenes by exploiting object-object qualitative spatial and s-t relations. Our system is capable of dynamically identifying reference objects and generates perceptual clusters. Then the system learns qualitative object-object spatial and s-t relations and thereby generates context specific rules that can be applied in any similar situation to learn the arrangement of the objects in the observed scene. According to the 70% of accuracy rate against a perfect scenario we can conclude that learning rules of a visual scene predominantly by analysing qualitative spatial and s-t relations is an effective way of learning rules of a visual scene.

At present we are researching how to incorporate rules learnt from one scenario for reasoning in

another scenario as humans do in real life situations. Our work can have great impact on building autonomous agents that learn from human agents based on visual inputs and can be used as a model for training agents even in a time of scarcity of human experts such as in a disaster.

8. Acknowledgement

The first author gratefully acknowledge the financial assistance under reference LKCN-2003-93 received from Commonwealth Scholarship Commission, United Kingdom, Professor A.G. Cohn for his guidance and supervision during the period at University of Leeds, the guidance and support given by the CogVis team at the University of Leeds and the CogVis team at the University of Hamburg, Germany for providing the data.

9. References

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

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

[3] N. I. Badler, Temporal Scene Analysis: Conceptual Descriptions of Object Movements, Technical report no. 80, University of Tronto, Ontario, Canada (1975).

[4] R. J. Howarth, H. Buxton,), "An Analogical Representation of Space and Time," Image and Vision Computing 10(7), 1992, pp. 467-478.

[5] J. Fernyhough, A. G. Cohn, D. C. Hogg, "Constructing Qualitative Event Models Automatically from Video Input," Image and Vision Computing, 18, 2000, pp. 81-103.

[6] B. Neumann, "Conceptual Framework for High Level Vision," Technical report FBI-HH-B-241/02, FB, Informatik, Univeritat Hamburg, 2002, June.

[7] D. Magee, C. Needham, P. Santos, A. G. Cohn, D. Hogg, "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.

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

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

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

[11] J. F. Allen, "Maintaining Knowledge about Temporal Intervals," Communications of the ACM 26 (11), 1983, pp. 832-843.

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

[13] E. Clementini, P. Di. Felice, D. Hernandez, "Qualitative Representation of Positional Information," Artificial Intelligence, 95, pp.317-356.

[14] E. Davis, "Representations of Commonsense Knowledge," editor, Brachman R.J., Morgan Kaufmann publishers, Inc, San Mateo, California, 1990.

[15] A. L. Kor, B. Bennett, "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.

[16] A. Galton, "Deriving Qualitative Spatial Information from Low- Level Data" unpublished.

[17] T. M. Mitchell, Machine Learning, Macgraw Hill Company. 1997.

[18] S. Muggleton, J. Firth, "Cprogol4.4; A Tutorial Introduction," In Relational Data Mining, editors, Dzeroski S., Lavarac N., Springer Verlag, 2001, pp. 160-188.

[19] D.D.M. Ranasinghe, A. S. Karunananda, 2006, "Qualitative Knowledge Driven Approach to Inductive Logic Programming," accepted for publication at ICIIS 2006.