Unsupervised Symbol Grounding and Cognitive Bootstrapping in Cognitive Vision

R.Bowden, L.Ellis, J.Kittler, M.Shevchenko and D.Windridge

Centre for Vision, Speech and Signal Processing,
The University of Surrey, Guildford, Surrey, GU2 7XH, UK

Abstract. In conventional computer vision systems symbol grounding is invariably established via supervised learning. We investigate unsupervised symbol grounding mechanisms that rely on perception action coupling[1]. The mechanisms involve unsupervised clustering of observed actions and percepts. Their association gives rise to behaviours that emulate human action. The capability of the system is demonstrated on the problem of mimicking shape puzzle solving. It is argued that the same mechanisms support unsupervised cognitive bootstrapping in cognitive vision.

1 Introduction

Computer Vision as a branch of Artificial Intelligence (AI) has, over recent years diverged from its roots. Forming a science within itself, it relies heavily upon statistical techniques. The recent interest in Cognitive Vision is, to some extent, an attempt to reunite computer vision with its AI roots.

Some argue that all computer vision is cognitive and the mere embodiment of a statistical technique, for example in robot vision, produces a cognitive system. Others actively try and combine modern day computer vision with more traditional AI approaches. However, the fusion of statistical and symbolic information raises a whole host of questions: What is the meaning of a symbol? How is it grounded? How is it constructed? Should symbols emerge or are the provided to the system a priori? Should symbols change/evolve? What syntax governs them and how can we learn this syntax irrespective of application? How can the system bootstrap itself to enhance its perceptual and reasoning capabilities?

The fundamental flaw in most cognitive vision development is its tailoring to specific applications. Transferable learning is a key aspect of cognitive approaches but a successful approach should provide generic learning, where, the same framework can be applied to a whole class of problems. This paper addresses some of the issues and presents early work towards generic symbolic processing and bootstrapping mechanisms for cognitive systems.

Following a discussion on symbol grounding in visual agents in Section 2, we present a novel unsupervised symbol grounding mechanism based on percept clustering and perception-action coupling in Section 3. Section 4 addresses the issues of cognitive bootstrapping. Finally conclusions are drawn in Section 5.

[1] This work was supported by EU projects COSPAL and VAMPIRE

2 Symbol Grounding

A central practical and philosophical problem for the construction of autonomous visual agents is that of *symbol grounding* [8]. An autonomous visual agent is, by definition, one capable of adapting to its in environment in behavioural and representational terms that go beyond those implied by its initial set of 'bootstrap' symbolic assumptions. Doing so necessitates the use of mechanisms of generalisation, inference and decision making in order to modify the initial perceptual symbol set in the light of novel forms of sensory data.

Any visual representation capable of abstract generalisation is implicitly governed by the laws of predicate logic. As such, the generalised entities must observe strictly formalised laws of interrelationship, and consequently, in abstracting the visual symbol set away from the original set of innate percept-behavioural pairings, are apt to become detached from any intrinsic meaning in relation to the agent's environment. A related difficulty, known as the *frame problem* [17], also arises in such generalised formal domains; it is by no mean clear which particular set of logical consequences (given the infinite number of possibilities) that the generalised reasoning system should concern itself with.

There is therefore a problem of symbol relevance and 'grounding' unless additional mechanisms can be put in place to form a bridge between the formal requirements of logical inference (applied to visual symbols), and the relevance of this symbol set to the agent within the context of both its goals and the intrinsic nature of the environment. In terms of the philosophy of cognition, this necessitates a move from a Quinean [9] to a Wittgensteinian [10] frame of reference, in which symbol *meaning* is intrinsically contextual, and environment-dependent.

For artificial cognitive agents embodied within the real world (that is to say, *robots*), the form that this symbol grounding framework must take is, by an increasing consensus (eg [11], [12], [13], [21]), one of hierarchical stages of abstraction that proceed from the 'bottom-up'. At the lowest level is thus the immediate relationship between percept and action; a change in what is perceived is primarily brought about by actions in the agent's motor-space. This hence limits visual learning to what is immediately relevant to the agent, and significantly reduces the quantity of data from which the agent must construct its symbol domain by virtue of the many-to-one mapping that exists between the pre-symbolic visual space and the intrinsic motor space [14]. For example, a hypothetical mobile robot engaged in simultaneous location and mapping (SLAM) (eg [16]) might build up a stationary stochastic model of any environmental changes that occur when not engaged in any direct motor activity, but switches to a Markovian transitional model when engaged in motor activity, thereby forming a sequence of 'key-frame' transitions driven by its motor impulses.

The first level of abstraction in the hierarchy thus represents a generalisation of the immediate, pre-symbolic percept-action relation into the *symbol domain*. There are many approaches to achieving this primary generalisation, for instance: unsupervised clustering [14], invariant subspace factoring [18], constructive solid geometry schematics [19]. Progressive levels of abstraction can be added by similar means, or they might instead involve higher levels of inferential machinery,

for instance first order logical induction for rule inference, if explicitly ascending the Chomsky hierarchy [15].

At some level of abstraction, critically, is the concept of *objects*, characterised by their persistence with respect to the agent's actions. Representations above this level are then characterised by their object-centric, rather than agent-centric descriptions (so we move from a percept-action space into a domain where descriptions with formal equivalents to English terms such as '*on*', '*under*', etc, can form part of the environment description). What results is a set of high-level, abstracted symbol generalisations that are nevertheless grounded in the percept-action space by virtue of the intermediate hierarchical levels. We might thus, for instance, envisage a tennis-playing robot that has the segmentation of the ball from the background at its lowest representative level, leading into a series of ascending representations that cumulate in the formal logical rules of the game of tennis at the most abstract level of representation. Furthermore, such a hierarchical structure has the advantage that higher-level action imperatives (such as, in our example, 'serving the ball') may act to reinforce learning at the lower-levels (by providing additional tennis-ball segmentation statistics). The hierarchical percept-action structure is hence robust and adaptive by nature.

A further possibility, once higher-levels of the perception-action hierarchy are sufficiently well established, is (by way of contrast to the previous passive example) to use these to *actively* drive lower-level learning. Hence, an inferred *partial* environment representation in an autonomous mobile robot might be used to initiate exploration into unmapped regions of the environment, or to improve upon weakly mapped environmental domains. Alternatively, percept clustering can itself be driven by higher-level concepts inferred from *those same* clusters, such as in Magee *et al.*'s [14] visual first-order logic induction system, in which clustered entities with identical inferred logical relations are deemed to be the *same* (which is to say they are meta-clustered by the higher-level inferential system).

Like the first example, this latter approach thus has the capacity to completely solve one of the critical difficulties of unsupervised cognitive learning that we have alluded to; the issue of *framing*. By deciding at what level to cluster entities in the sensory domain on the basis of entities formed from those same clusters, the potential for redundant higher-level inferences from the sensory domain can be entirely eliminated. The symbol structure thus becomes *entirely* agent-relative, irrespective of the initial set of symbolic assumptions with which the system was 'bootstrapped' into cognitive activity. We hence term this active hierarchical-feedback approach to autonomous cognition **'cognitive bootstrapping'** by virtue of the capacity of the cognitive systems so described to make their symbolic representations fully *self-foundational*.

It is apparent that classical AI approaches to cognitive vision were unsuccessful in that they attempted to build a high-level environmental description *directly from* the percept space before going on to consider agent actions within this model, rather than allowing this representation to evolve at a higher hierarchical level [20]. Representative priorities were thus specified in advance by the

system-builder and not by the agent, meaning that *autonomous* agency had to build its goals and higher-level representations in terms of the *a priori* representation, with all the redundancy that this implied. Furthermore, novel modes of representation were frequently ruled out in advance by the pre-specification of scene-description.

In the following sections we shall discuss a few mechanisms that accomplish symbol grounding without conventional learning in a supervised mode. The symbol grounding is achieved by associating percepts with actions. This association gives rise to interesting perception - action behaviours. As an example, in the next section, we demonstrate that the system can learn to play a game such as puzzle. More interestingly, we show that unsupervised clustering and/or quantisation of percepts and observed actions lead to the discovery of new concepts and functionalities, characteristic of bootstrap learning and emergence of intelligence.

3 Modelling Perception Action Coupling

In this section we present a framework for autonomous behaviour in vision based artificial cognitive systems by imitation through coupled percept-action (stimulus and response) exemplars.

The assumption is made that if a cognitive system has, stored in its memory, a representation of all the possible symbolic perceptual stimuli (percepts) that it shall ever encounter each coupled with a 'symbolic' action model (response), then it should be capable of responding as required to any given visual stimulus. This assumption leads us to consider how a practical estimation to such a system could be achieved.

The system must be capable of searching its entire percept store (visual memory) in order to find a match to the current percept (visual stimulus), and then perform the associated action.

By hierarchically grouping the percepts into progressively more general representations(expressions), and given that the stored percepts adequately cover the percept space, we can structure the stored percepts in such a way as to allow fast searching of the percept space. Also in generalising the percept representations further at each level of the hierarchy, the system is capable of compensating for the incomplete coverage of the percept space by performing more general action models given more general percept representations.

This system operates within a simulated shape sorter puzzle environment. The training phase is initiated by the supervisor solving the shape sorter puzzle a number of times. Each time the supervisor takes some action, the system records both the action performed and the associated visual percept. During the systems on-line state it extracts the current scene information, builds a symbolic representation in order to compare to the stored percepts and then performs the action associated with the best matching percept.

In order to cluster percepts at each level of the hierarchy, the representation must allow us to measure similarity between two percepts. In this system a scene

is represented symbolically as an Attributed Relational Graph. Graph vertices represent the objects in the scene. Graph edges represent the relational structure of the scene, see figure-1. *Type* attributes are attached to each vertex and dictate the type of object. Graph edge attributes are 3D *relative_position/orientation* vectors that represent both the horizontal and vertical displacement and the relative orientation between the two objects connected by the edge.

Formally we define Attributed Relational Graphs (ARGs) as a 4-tuple, g = (V,E,u,v) where V and E (E $\subseteq$ V x V) are the set of nodes (graph vertexes) and edges (links between nodes) respectively. u : V $\longrightarrow$ Av is a function assigning attributes to nodes, and v : E $\longrightarrow$ Ae is a function assigning attributes to edges. Av and Ae are the sets of node and edge attributes respectively.

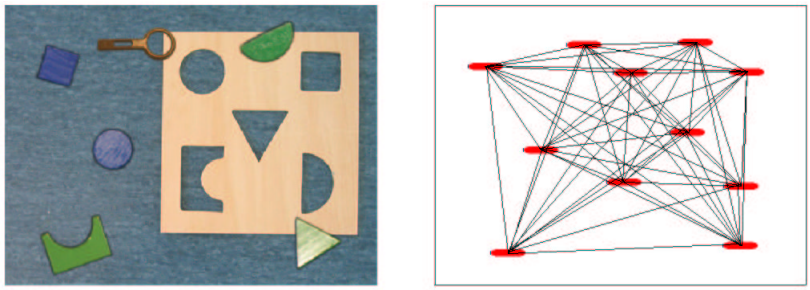

Fig. 1. Percepts are represented as Attributed Relational Graphs.

In order to group the percepts we need some way to measure/compute the similarity or distance between two Attributed Relational Graphs. We have adopted the VF graph matching algorithm [7] developed at the Artificial Vision Group (University of Naples). The median of each group/cluster is computed and is used to represent the cluster members at the current level. The median graph is computed by finding the graph that has the minimum sum of distances to all other cluster members.

In order for a cognitive system to actively respond to visual stimulus, a mapping between percepts and actions is required [1]. Recent neurophysiological research has shown strong evidence supporting the existence of a mechanism, in both primate and human brains, that obtains this percept-action coupling, known in lit. as "direct-matching hypothesis". The mirror-neuron system essentially provides the system (human/primate brain) with the capability "to recognise actions performed by others by mapping the observed action on his/her own motor representation of the observed action" [2].The system presented here has been endowed with these same innate capabilities. Therefore our work differs from that of [3] where an attempt is made to analyse from visual data, the force dynamics of a sequence and hence deduce the action performed. Instead, by allowing the system to directly record symbolic representations of the actions performed during the training phase, an exact percept-action coupling can be achieved. As an alternative approach, [4] has shown that it is possible for an agent to learn to mimic a human supervisor by first observing simple tasks and

then, through experimentation, learning to perform the actions that make up the tasks.

h

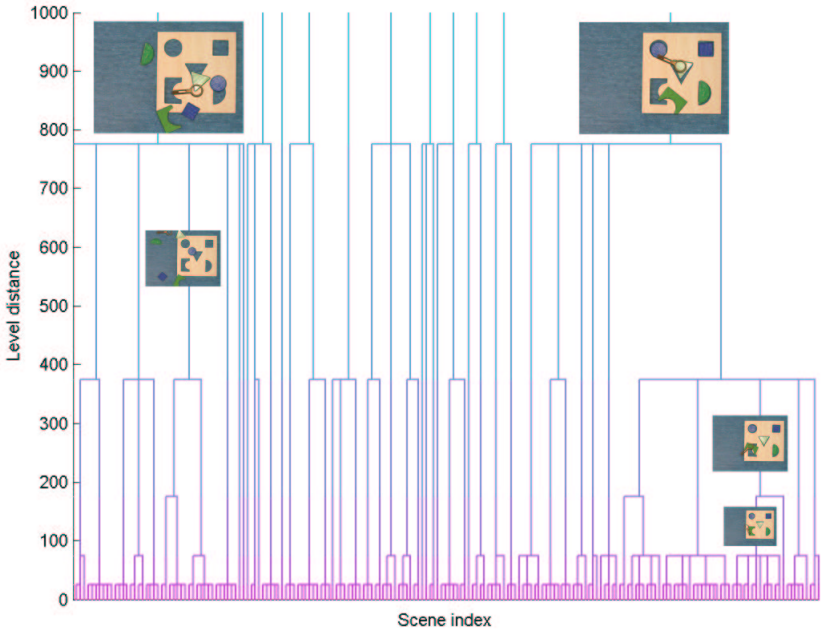

Fig. 2. Percepts are clustered into a hierarchy.

The lowest level actions that the system is capable of performing are 'pick up object', 'put down object' and 'move gripper'. Within the context of the problem the system needs only to perform these actions in a fixed format: A:= move - pick up - move - put down. Since some of the intermediary percept-action pairs are no longer needed (e.g. those percepts coupled to the 'pick up' or 'put down' action), *key scenes* corresponding to the beginning of a sequence, must be extracted and coupled with the action model. It is these key scenes, represented as ARGs, that form the data set of the systems visual memory. It is worth making clear here that we have temporally segmented the continuous perceptual input in accordance with the beginning of our fixed format action sequences and discarded all the perceptual data not extracted as key scenes. This approach has the advantage of reducing the amount of data needed to store an entire puzzle sequence.

Another motivation for fixing the action format is that a fixed length vector can now be used to represent an action. To model the actions that are coupled to the key scenes, only a five-element vector is required. The first and second elements in the action vector represent the change, brought about by the first 'move'

action, in the horizontal and vertical positions respectively. As the 'pick up' and 'put down' parts of the action need no parameterisation and are implicit to the action, they are not represented in the action vector. The last three elements of the action vector are therefore left to parameterise the final 'move' operation. This is the same as the first 'move' but with a third 'rotation' dimension.

Now that an action can be modelled as a vector, the actions coupled to a cluster of percepts can be represented as a matrix. Note that the action clusters are a result of perceptual grouping.

Many of the activities that cognitive systems must perform require perceptually guided action e.g. walking or catching. It is also true to say that much of what is perceived by a cognitive system is the result of actions performed by that system e.g. moving to a new space in the environment or picking up an object to examine it. The perception-action cycle simply describes a model of behaviour whereby perception influences action and action influences perception.

The system extracts a symbolic representation, percept, from the current sensory input. It then finds the closest matching percepts in its visual-proprioceptive memory by searching the percept hierarchy. The resulting matches are each associated with an action vector and, depending on the level in the hierarchy at which the match is made, a generality parameter. The action is performed and so the scene changes and once again the system extracts a new percept and so the cycle continues until a solution is reached. The operator must *teach* the system by solving the puzzle a number of times. This provides the system with the expected behaviour that it will attempt to imitate during game play.

4 Cognitive Bootstrapping

The framework described in the previous section endows the system with the ability to imitate action behaviours. [5] have recently presented a framework for learning of object, event and protocol models from audio visual data. Their framework employs both statistical learning methods, for object learning, and symbolic learning for sequences of events representing implicit temporal protocols. As [5] point out this learning of temporal protocols is analogous to grammar learning, in this respect the system presented here shares some goals with that presented by [5]. Further, both systems attempt to achieve this grammar learning through generalising a symbolic data set. There is however very little similarity between the approach taken by [5]; Inductive Logic Programming, and that which we have taken; clustered percept-action exemplars. Where [5] have developed, using Progol, an inference engine to extract temporal protocols, we have employed an *approximation to imitation* approach to learning puzzle grammars/temporal protocols.

It should be noted that we have given the system no indication as to what the goal of the activity is and what represents a successful conclusion of the game. In fact without any other functionalities the system would be unable to change its behaviour and reason about unusual states. To correct this, the system needs a mechanism that can discover the goal of an observed activity. Once the goal is

defined, the behaviour can be optimised or even adapted by direct optimisation of the objective function that encapsulates the activity goal. The goal can be discovered by clustering observed actions or by clustering the perceptual stimuli corresponding to the activity end state. The *solved puzzle* state is recorded by the system each time the user finishes the puzzle. Such perceptual data analysis would show the end state to be unique and reaching it must be the goal of the behaviour. The definition of the goal then allows a cost function to be set up in terms of distance to end state.

In order to allow the system to improve its performance at a given task over time, during its own interactions with the world, it stores the resulting percept-action pairs. The system is capable of supervising itself over time by rewarding near optimal action sequences. Optimality is defined in terms of the distance between the current percept and a percept relating to a *solved puzzle* state.

In addition to self optimisation, the existence of the goal leads to a new type of behaviour where at each step of the game an action minimising the distance to the end state is selected from the list of available actions. The monotonic behaviour of the cost function for a sequence of actions leading to a solution is shown in figure 3. An equally effective solution could be found by clustering actions. This *hard wired* ability to explore the action space in the direction of its modes or in the directions of the modes of the percept space is considered to be innate. It is instrumental in providing bootstrapping capability, as it enables the system to discover new solutions and to reason about the task.

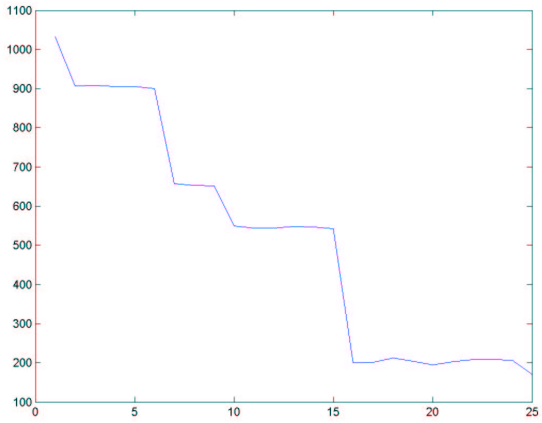

Fig. 3. The cost function plot for a successful sequence of actions leading to a solution.

At this point it is pertinent to ask, how the vision system could bootstrap itself when it is first switched on and has no prior perceptual capabilities. The above shape puzzle problem has been solved under the assumption that the sys-

tem can segment objects by colour and shape and is able to move its arm/gripper intentionally from point to point, as well as pick and place objects. Unless one opts for hard wiring, such functionalities have to be acquired by self learning.

We have studied this problem and shown that perception action cycle, combined with the unsupervised learning and action space exploration mechanisms discussed above are sufficient to build up vision capabilities from the ground level. The only assumption we make is that the system is able to drive its arm, in a random manner and observe its movement. This has been simulated by representing the arm as a single point in a 2D work space and moving it from one point to another. The random moves were constrained by hard environmental boundaries. Such constraints are realistic, as they emulate the world being dominantly horizontal, subject to the laws of gravity. Thus random motions will often be confined to horizontal and vertical directions imposed by the constraints.

Unsupervised clustering of the observed visual data detect these dominant directions which then activate the exploration module to attempt arm movements in the required directions. As the system, initially, has no link between action and perception, the dominant directions in the percept space provide an action goal the achievement of which can be measured in the percept space. This internal goal of the system and the observed error allow the system to learn to perform the intentional actions effectively in a supervised mode.

Once the system is self trained to move its arm in the two basic directions, the system discovers that driving the arm will result in its displacement and that the distance travelled horizontally or vertically will depend on the strength of the driving force (frequency and duration of the motor neuron signal). Also, the system is now able to move from point A to point B in a city block manner. By self optimising, the system eventually acquires the ability to move intentionally from any point in its work space to any other and to establish a link between perception and action. The acquisition of other low level vision and action capabilities in a bootstrapping mode is currently investigated and will be demonstrated by the system in the future.

5 Conclusions

We addressed the problem of symbol grounding in cognitive vision. In conventional computer vision systems symbol grounding is invariably established via supervised learning. However, such approaches make the system too application specific and provide no capability for self learning, self optimisation, acquisition of novel behaviours and general reasoning. We investigated unsupervised symbol grounding mechanisms that relied on perception action coupling. The mechanisms involved unsupervised clustering of observed actions and percepts. Their association gave rise to behaviours that emulated human action. The capability of the system was demonstrated on the problem of mimicking shape puzzle solving. We showed that the same mechanisms supported unsupervised cognitive bootstrapping in cognitive vision.

References

1. Granlund, G. 2003. Organization of Architectures for Cognitive Vision Systems. In *Proceedings of Workshop on Cognitive Vision.*
2. Buccino, G.; Binkofski, F.; and Riggio L. 2004. The mirror neuron system and action recognition. In *Brain and Language, volume 89, issue 2, 370-376.*
3. J. M. Siskind 2003. Reconstructing force-dynamic models from video sequences. In *Artificial Intelligence archive, Volume 151 , Issue 1-2 91 - 154.*
4. P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini. Learning About Objects Through Action: Initial Steps Towards Artificial Cognition. In *2003 IEEE International Conference on Robotics and Automation (ICRA).*
5. Magee D., Needham C.J., Santos P., Cohn A.G. and Hogg D.C. Autonomous learning for a cognitive agent using continuous models and inductive logic programming from audio-visual input. In *Proc. AAAI Workshop on Anchoring Symbols to Sensor Data, 17-24.*
6. Nock R. and Nielsen F. Statistical Region Merging. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 26, Issue 11, 1452- 1458*
7. Cordella L., Foggia P., Sansone C., Vento M. An ecient algorithm for the inexact matching of arg graphs using a contextual transformational model. In *Proceedings of the International Conference on Pattern Recognition, volume 3, 180-184*
8. Harnad, S. (1990). The symbol grounding problem. *Physica D, 42,* 335-346.
9. Quine, W. von O., 1960. Word and Object. New York: John Wiley and Sons, Cambridge: MIT.
10. Wittgenstein, L., Anscombe, G. E. M. (translator), 'Philosophical investigations : the German text with a revised English translation by Ludwig Wittgenstein', Oxford : Blackwell, 2001, ISBN 0631231277.
11. Marr, D. (1982). Vision. San Francisco: Freeman.
12. Gärdenfors, P. 1994. How logic emerges from the dynamics of information. In Van Eijck/Visser, Logic and Information Flow, 49-77.
13. Granlund G., Organization of Architectures for Cognitive Vision Systems, 2003, Proceedings of Workshop on Cognitive Vision, Schloss Dagstuhl, Germany.
14. D. Magee, C. J. Needham, P. Santos, A. G. Cohn, and D. C. Hogg (2004), Autonomous learning for a cognitive agent using continuous models and inductive logic programming from audio-visual input, AAAI Workshop on Anchoring Symbols to Sensor Data.
15. Chomsky, N., Three models for the description of language, IRE Transactions on Information Theory, 2 (1956), pages 113-124
16. Thrun S.,2002, Robotic Mapping: A Survey, Exploring Artificial Intelligence in the New Millennium. Morgan Kauffmann
17. McCarthy, J. & Hayes, P.J. (1969), Some Philosophical Problems from the Standpoint of Artificial Intelligence, in Machine Intelligence 4, ed. D.Michie and B.Meltzer, Edinburgh: Edinburgh University Press, pp. 463-502.
18. Granlund G. H. and Moe A., Unrestricted Recognition of 3D Objects for Robotics Using Multilevel Triplet Invariants., AI Magazine, vol .25, num. 2, 2004, p51-67
19. A. Chella, M. Frixione, S. Gaglio: A Cognitive Architecture for Artificial Vision, Artif. Intell. 89, No. 1-2, pp. 73-111, 1997.
20. Brooks, R. A., Intelligence without Representation, Artificial Intelligence, Vol.47, 1991, pp.139-159
21. Johnson-Laird P. N., Mental Models, Harvard University Press, Cambridge, MA 1983.