Cognitive Vision

Background and Research Issues

Gösta Granlund
Computer Vision Laboratory
Linköping University

November 2002

1. Introduction

Systems for handling and understanding of cognitive information are expected to have as great impact on society over the next decades, as what conventional computers and telecommunication have on today's society. They promise to relieve humans of many burdens in the use and the communication with increasingly complex systems, be they technical or deriving from an increasingly complex society. They will make many new applications possible, ranging from autonomous home appliances to intelligent assistants keeping track of the operations in an office.

Up until now, systems have been built, which can operate in very restricted domains or in carefully controlled environments – i.e. in artificially constrained worlds – where models can be constructed with sufficient accuracy to allow algorithms to perform well. However, we want systems that can respond to and act in the real world. The real world is very complex, and there is no possibility to specify all alternative actions and the decision criteria for these in the traditional way.

Cognitive systems need to acquire the information about the external world through *learning* or *association,* as the complex interrelationships between percepts and their contextual frames could never be specified explicitly through programming. The fundamental mode of operation for learning is that *action precedes perception*. This is because the action space is much less complex than the percept space, and can drive an association process. This implies that cognitive systems have to be developed in a full perception-action feedback cycle.

The functional context is important, as we rarely process information in an intentional vacuum – we always have goals. Representation of context at lower levels of a cognitive system is more complex and spatial/quantitative than we are used to for linguistic descriptions. Linguistic descriptions require the mapping into spatial-perceptual parts of a cognitive system, where references are made to its own acquired spatial knowledge and the actual state of the system.

2. Overview of Research Issues in Cognitive systems

The purpose of cognitive systems is to produce a response to appropriate percepts. The response may be a direct physical *action* into the environment of the system. Such an action will somehow change the *state* of the system, which allows us to interchangeably say that percepts shall be related to actions or to states. The response may be delayed in the form of a reconfiguration of internal models in response to the interpreted *context* of the system. Or it may be to generate in a subsequent step a generalized *symbolic representation,* which will allow its intentions of actions to be communicated.

There is some debate as to what exactly constitutes cognitive systems – especially where they start and where they end. Several terms such as perception, cognitive systems, AI, etc., may in different cultures represent partially or totally overlapping concepts, while they in others take on very specific connotations. Rather than trying to make some unambiguous definition, this document will propose areas of research which will contribute to a common goal of devising systems which can perceive and learn important information in an interaction with the environment and generate appropriate, robust actions or symbolic communication to other systems, e.g. in the form of language to humans. This defines the use of the term cognitive vision for the purpose of this document.

The inputs to a cognitive system, or the representations of information in early stages of it, are generally referred to as *percepts*. They will typically be visual or auditory, as these modalities generally carry most information about the environment. However, other sensing modalities may be used, in particular for bootstrapping or other support purposes. Perception and percepts are similarly ambiguous terms, where some may say that perception is in fact the function performed by a cognitive system. However, there is generally agreement that percepts are compact, partially invariant entities representing the sensing space in question. Visual percepts will for example be some processed, more invariant, more compact representation of the information in an image, than the original iconic image obtained from the sensor.

Much of the lack of success in vision for complex problems can be traced to the early view that percepts should generate a *description* of the object or the scene in question. This description has typically been in geometric terms with a conceptual similarity to CAD representations. See Figure 1. This description should then be used to implement *actions*.

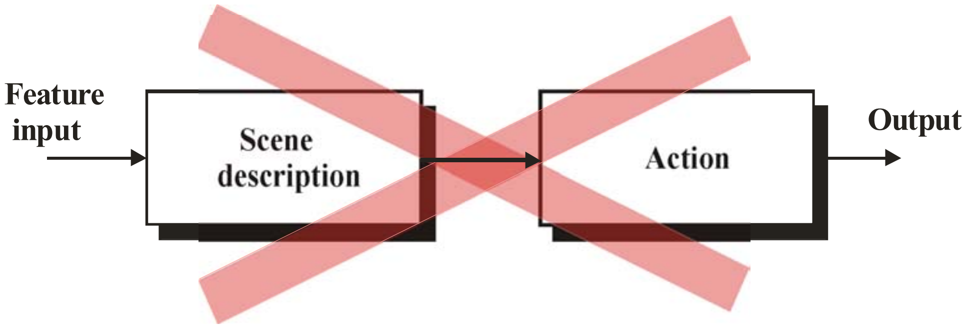

Figure 1

The problem with a description of an object or a scene is that it lacks an *interpretation*, i.e. links to actions that are related to the object. The purpose of a Cognitive Vision system is to build up a model structure which relates the percepts emerging from an object to its states, or actions performed upon it. This allows the system to separate an object from its background, separate distinct parts within an object, learn how the percepts transform under manipulation, etc. Actions can likewise be used to manipulate the environment, which in consequence will modify the emergent percepts. Learning these relations gives the system the information required for the subsequent use in the opposite direction: To use percepts to control actions in a flexible fashion.

A description, on the other hand, is a representation at the symbolic level of a cognitive system, which is to be used for planning and for communication. A description in the perception-action part of the structure is at worst an intermediary step, leaving too little information for a subsequent actuator to perform proper actions. We will deal with these aspects later.

Most researchers today agree that the order of the processes should rather be the opposite. See Figure 2. A cognitive system should start out with a reactive percept-to-action mapping process. There are a number of additional reasons for this. One reason is that this allows a feedback piece-wise continuous structure of a type required for an exploratory learning process, necessary to derive sufficiently detailed models of the system's environment. Another reason is that *learning is driven from the action side* of a cognitive system rather than the percept side, because the state complexity of the action side is much lower. In consequence, this allows the system to find the percepts which are likely to be involved in the generation of a certain action. The strategy is using what is termed *view-centered* object representation.

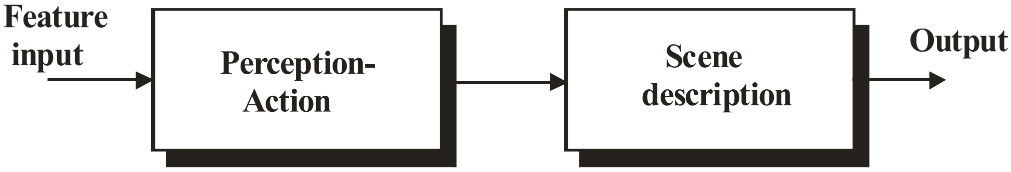

Figure 2.

It also allows exploratory strategies for learning, where it uses earlier acquired information to make the process more efficient. *Active vision* is another term for such exploratory strategies. The preceding is also the strategy which is believed to be used by biological systems. For the purpose of this document, we will simply refer to this part as the one performing *perceptual* processing or being the *perceptual* domain. This is in contrast to the later discussed *symbolic* processing. A more detailed structure is given in Figure 3.

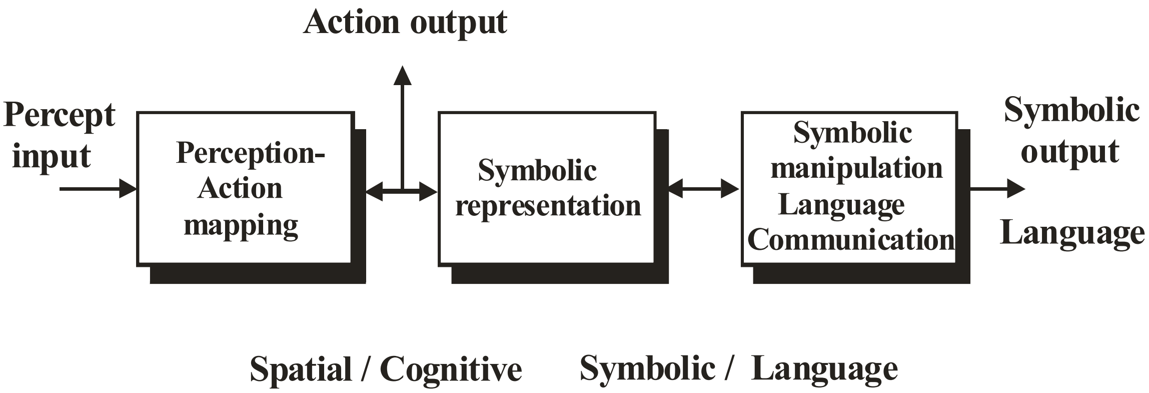

Figure 3.

It is believed that the subsequent symbolic representation shall emerge from, and be organized around, the action or state representation, rather than from any descriptive, geometric representation. This does not exclude the use of static clues such as color.

There are strong indications that this is the way it is done in biological systems --- it is known that our conscious perception of the external world is in terms of the actions we can perform with respect to it. From an evolutionary view, lower level organisms essentially only have the perception-action mapping, while the descriptive, symbolic representation is a later development, though extremely important. The main argument for this strategy is, however, that it gives a realistic path for development of evolutionary/learning technical systems, ranging from low percept levels to high symbolic reasoning levels.

The transition from the action or state representation to a symbolic representation, implies in principle a stripping off, of detailed spatial context to produce sufficiently invariant packets of information to be handled symbolically or to be communicated. This may require the introduction of symbolic contextual entities, derived from certain contextual attributes in the perceptual domain. What has earlier been termed a description is equivalent to a symbolic representation. This is also the part of the system where descriptions such as categories of objects should emerge.

Subsequently follows the symbolic processing structure with its different implementations, such as for planning, language and communication. Symbolic representation and manipulation should be viewed as a domain for efficient processing of concepts in a relatively invariant format without unnecessary spatial, contextual qualifiers, which would severely complicate the processing. The invariant format makes manipulation and communication much more effective and its effects more generally applicable. While the perception-action structure deals with *here-and-now*, the symbolic structure allows the system to deal with other points in space and time in an efficient way. This is also what allows *generalization*. A symbolic representation is on the other hand a too meager form for sufficiently adaptive control of actions, a sometimes overlooked characteristic of language. Language works in spite of its relative low information content, because it maps onto a rich spatial knowledge structure at all levels, available within our surprisingly similar brains.

The output from a symbolic representation and manipulation is preferably viewed as designed for *communication.* This communication can be to another system, or to the perceptual processing part of the *own* system. This implies that the symbol structure is converted back to affect a fairly complex and detailed percept-to-action structure, where contextual and action parameters are reinserted in a way related to the actual state of the system. In this way, symbolic information can be made to control the perception-action structure, by changing its context. The change of context may be *overt* or physical in commanding a different state or position bringing in other percepts, or *covert* affecting the interpretation of percepts. The symbolic representation must consequently be translatable back to detailed contextual parameters relating to the actual state of the system, be it the own system or another system.

A significant trend over the last years is the recognition of the importance of *semantics* compared to *syntax*. This implies that many important relations between entities can not be described and predicted by rules. These relations simply appear as *coincidences* with no simple predictability models. Still, they are extremely important for systems intended for the real world, as the real world has this nasty unruly habit. The only way to acquire this information is through *association* or learning. This is true all through a cognitive system, from percepts to language, while the implementation will be different for different parts.

The preceding strongly emphasizes the necessity for a full fledged cognitive system to have both a spatial/perceptual part and a symbolic/language part in close integration. An important issue is that the spatial/perceptual domain and the symbolic/language domain are two different worlds, where different rules and methods apply. This includes how information is represented, how learning is implemented and consequently how memory is represented.

3. Use of cognitive mechanisms for development of vision systems

In most of the preceding discussion, a feedback perception-action structure has been assumed, which primarily reminds of robotics applications. Does that mean that the preceding methodological structure is only applicable to robotics?

No! The belief is that the structure discussed is not only advantageous, but necessary, for demanding applications of vision including static imagery. Similarly for cases where the output is not a physical action but the communication of a message. An example of the latter type is man-machine interfaces, where the actions and speech of a human are registered, interpreted and communicated symbolically to a system to implement a very sophisticated control of its functions. Sufficient flexibility and adaptation requires learning for the system to deal with all contextual variations encountered in practical situations.

The training of cognitive systems for such advanced but non-robotic applications requires the development of mixed real-virtual training environments. In these, the system will gradually build up its knowledge of its environment with objects including humans. The learning is again implemented as *association*, between the learning system's own state parameters, and the impinging perceptual parameters. The typical case is as discussed earlier that the system moves an object in front of its camera input. The movement parameters are known to the system and can be associated with the percepts appearing as results of the movements. This can in training environments be simulated in various ways such that corresponding state and percept information is made available to the system.

In such a way, competence can be built up step by step, in a mixture of real and virtual training environments. With a design allowing incremental learning, it shall be possible to start out with reasonably crude virtual environments, to give the system some tentative knowledge of object and environment space structure, which is refined in the real environment. An important feature is that *copies* can be made of a trained system, to allow an efficient production of systems. On the other hand, it may not be possible to easily copy particular information from one trained system to a differently trained system. This is because new information has to be incorporated on the terms of the system acquiring it, i.e. connected to other stored information. This can not be implemented as a copying process, but the information can be supplied over the normal channels, where corresponding state and percept information can be made available to the system for organization on its own terms.

From this derives the view that the development of powerful cognitive systems inevitably has to go the path over perception-action mapping and learning, similarly to the case for robotics, even if the systems will be used for interpretation of static imagery or to generate and communicate messages to other systems. It should be noted that the perception-action mapping is in fact implementing a training of the system to point out what is essential for its understanding of the external world, but a training which the system to a substantial part can implement on its own.

This opens up wide ranges of applications of cognitive systems at an early stage, which do not require advanced mechanical manipulators. One such important application field is in activity interpretation for man-machine interfaces. Setting up training environments for such applications is one of the research challenges.

4. Future Research and Development

Research in Cognitive Vision should be carried out in a systems perspective. While all research can not be expected to cover complete systems, there should be an awareness of how research fits into a more complete systems perspective. A major objective is the development of systems or mechanisms allowing performance to be *extendable*, using learning, adaptivity, etc., rather than implementing “canned” predefined tasks. A goal is to develop systems which can operate in relatively unrestricted and changing environments. Ultimately a key issue is to achieve *behavioral plasticity*, i.e. the ability of an embodied system to learn to do a task it was not explicitly designed for.

As this is interdisciplinary research, inspiration as well as active participation is expected from fields such as systems and computer science, perceptual psychology, neurobiology, linguistics.

Important research issues:

- Better learning structures having larger capacity and permitting faster training, while still in batch mode. Such structures may employ new types of information representation having e.g. sparseness for efficiency and locality for faster convergence in training.

- Structures for incremental learning. This is ultimately required for advanced applications in the future, to allow a cognitive system to continuously adapt in real time to tasks and environment. This introduces even stricter demands upon choice of information representation and implementation of memory, such that just acquired information can be immediately related to stored information.

- Development of sparse, efficient feature sets for use in learning structures.

- Efficient mechanisms for learning perception-to-action mapping. This will make severe requirements upon how information is represented and require high capacity learning structures using this representation.

- Development of learnable model structures for representation of object parts – objects - relations - scenes in a common framework. For a model structure to be extendable, it is necessary that acquired lower level models or primitives can be reused as parts of new higher level models. The current view is that a low-to-medium level model structure is preferably expressed in terms of perception-to-action relations.

- Development of structures to represent context. This shall ultimately be integrated with the previous item, as it forms the structural "glue" or the reference structure between entities.

- Structures for transition from spatial/perceptual representation to symbolic/language representation. This implies in principle stripping off detailed spatial contextual parameters. Choice of information representations on both sides is crucial, such that e.g. confidence or certainty can be carried along between the domains.

- Structures for reinsertion of spatial/perceptual context from the symbolic/language domain, in relation to actual state of the system. This research should preferably be done in parallel with the previous item, as they are obviously strongly related and require the same background of concepts.

- Symbolic processing structures using new information representations which can better handle *similarity* and *metric* to build up *semantic spaces*. This includes use of confidence measures propagated from the spatial/perceptual processing part.

- Development of structures for acquisition or learning of symbolic semantic models within the framework of the previous item. The present item is different from the preceding model learning structure for learning of spatial/perceptual models, with respect to its organization, information representations, etc.

- Management of complexity using distributed control in cognitive systems:
    - Balance between centralized and distributed control.
    - Information representations allowing adaptivity, to establish proper connection and communication between system parts
    - Obtaining a coherent global behavior from the adaptive interaction of distributed system parts, not knowing each other's functionality but only certain input-output characteristics.
    - How to structure a system to avoid the convergence fallacy. (The grandmother cell)

- Representation of memory is an important issue in many of the preceding items. The implementation is anticipated to be different for the spatial/perceptual part and the symbolic/language parts. It is directly related to the representation of information used, and it may e.g. turn out that a distributed organization is advantageous for parts of the system.

- Development of strategies and implementations for mixed real-virtual training environments for cognitive systems.