Organization of Architectures for Cognitive Vision Systems

Goesta H. Granlund

Computer Vision Laboratory, Linkoeping University
581 83 Linkoeping, Sweden
gegran@isy.liu.se,
WWW home page: http://www.isy.liu.se/gosta

March 26, 2004

Abstract. The purpose of cognitive systems is to produce a response to appropriate percepts. The response may be a direct physical *action* which may change the state of the system. It may be delayed in the form of a reconfiguration of internal models in relation 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 to some other system. As important as the percepts, is the dependence upon context.

A fundamental property of cognitive vision systems is that they shall be *extendable.* This requires that systems both acquire and store information about the environment autonomously – on their own terms. The distributed organization foreseen for processing and for memory to allow learning, implies that later acquired information has to be stored in relation to earlier. The semantic character of the information which this requires, implies a storage with respect to similarity, and the availability of a metric.

The paper discusses organization aspects of such systems, and proposes an architecture for cognitive systems. In this architecture, which consists of two major parts, the first part of the system, step by step performs a mapping from percepts onto actions or states. The central mechanism is the *perception-action* feedback cycle. In a learning phase, *action precedes perception.* The reason for this reversed causal direction is that action space is much less complex than percept space. It is easy to distinguish which percepts change as a function of an action or a state change. Percepts shall be mapped directly onto states or responses or functions involving these.

The second part of the architecture deals with more invariant representations, of *symbolic* representations, which are derived mainly from system states and action states.

Through active exploration of the environment, i.e. using perception-action learning, a system builds up concept spaces, defining the phenomena it can deal with. Information can subsequently be acquired by the system within these concept spaces without interaction, by extrapolation using passive observation or communication such as language.

This structure has been implemented for the learning of object properties and view parameters in a fairly unrestricted setting, to be used for subsequent recognition purposes.

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 todays 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 also need 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, by supplying information in some declarative form.

Cognitive systems need to acquire the information about the external world through exploratory *learning*, as the complex interrelationships between percepts and their contextual frames can not be specified explicitly through programming with any reasonable efforts.

In the subsequent discussion, there will be several references to known properties of biological vision systems. It should be emphasized at the outset that the ambition of this paper is not to argue possible models of biological vision systems, but to propose potentially effective architectures of technical systems. In this process, however, it is deemed useful to take hints from what is known about biological vision systems.

2 Characteristics of Cognitive Systems

The purpose of cognitive systems is to produce a response to appropriate percepts. The response may be a direct physical *action* which may change the state of the system. It may be delayed in the form of a reconfiguration of internal models in relation 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. As important as the percepts, is the dependence upon context.

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 human language. This defines the use of the term cognitive vision in 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 boot-strapping 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.

A fundamental property of cognitive vision systems is the *extendability.* This implies that a system shall be able to deal with more situations than exactly those which the designer has foreseen, and programmed it for. This requires that systems both acquire and store information about the environment autonomously – on their own terms. The distributed organization foreseen for processing and for memory to allow learning, implies that later acquired information has to be stored in relation to earlier. The semantic character of the information implies a storage with respect to similarity, and the availability of a metric.

Building a complete vision system is a very difficult task. This difficulty results from:

- The huge amount of data to treat and the necessarily associated drastic information compression
- The necessity to combine goal driven and event driven inference mechanisms
- The necessity to design systems in a fragmented or multi-level structure

It has become increasingly apparent that classical computing architectures designed for symbol strings, are not appropriate for the processing of spatial-cognitive information. One reason is that the inevitable requirement of learning does not go well with the traditional separation of memory from processing resources. In the task to develop efficient architectures for technical vision systems, it is tempting to look at architectures of biological systems for inspirations on design [20, 3].

The views on cognitive vision architectures range between two extreme views [5]:

- Knowledge and scene representations must be supplied by the designer to the extent possible
- Knowledge about the external world has to be derived by the system through its own exploration

Proponents of the first view argue that if working systems are going to be available in some reasonable time, it is necessary to supply available information, and the modality for this is in declarative form. Proponents of the second view argue that if sufficiently complex models of the external world are going to be available to the system, it is necessary that it can explore and find these out essentially by itself.

We can assume that the economically optimal variety at the present time is somewhere between these extremes, and some combination of these. The present paper will adhere to the second view above.

The difficulty is not just to find solutions for identified sub-problems but to find a coherent, extendable and efficient architecture to combine the required functionalities. The issue is what strategies can be shown to be advantageous for different parts of cognitive systems, e.g. for the commonly assumed distinctive parts for relatively continuous perception-action mapping, versus discrete symbolic processing. This will also involve choice of preferable information representations of information and generic classes of operations, representation of memory in relation to different modes of learning, representation of uncertainty in such structurally different parts of a cognitive system.

Various alternatives for cognitive organization have been studied. One such approach is Bayesian inference and modeling [15]. Other issues are to explore particular types of information representations, which may have advantages in allowing mode separations in large systems. One such representation is the channel representation [6], which allows a fast convergence of learning structures, due to its locality property.

3 Overall Architectural Considerations

A vision system receives a continuous barrage of input signals. It is clear that a system cannot attempt to relate every signal to all other signals. What mechanisms make it possible to select a suitable subset for processing, which will not drown the system?

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. There has traditionally been a belief that an abstraction of objects should be generated as an intermediary step, before a response is synthesized. A great deal of the robotics field has been devoted to the generation of such generalized descriptions [12]. The classical model for image analysis in robotics has been to build up a description step by step. This description has typically been in geometric terms with a conceptual similarity to CAD representations. See Figure 1. The ambition has usually been that the model should describe the object geometry as accurately as possible. An extreme but common form of this is the statement that *...the best model of an object is the object itself.*

Given this description of the image, objects shall be recognized and assigned to the proper categories, together with information about position and other relevant parameters. This description is then carried to a second unit where it is interpreted to generate appropriate actions into the physical world, e.g. to implement a robot.

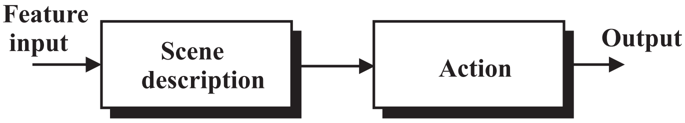

Fig. 1. Classical robotics model

This structure has not worked out very well for reasons which we will deal with in subsequent sections. In brief, it is because the leap between the abstract description and the action implementation is too large. A large number of important contextual qualifiers, of spatial and temporal nature, necessary for precise action have been lost in the abstraction process implicit in a description.

The major problem with this structure is that we primarily do not need a *description* of an object or a scene. What we need is an *interpretation*, i.e. links between actions and states that are related to an object and corresponding changes of percepts. The purpose of cognitive vision systems is consequently not primarily to build up models of the geometry of objects or of scenes. It is rather to build up model structures which relate the percept domain and the action or state domain; to associate percepts emerging from an object, to states or actions performed upon the object.

To achieve this, it turns out to be necessary to break up the big leap between percept structure and action structure into a sequence of smaller steps. In each such limited step, percepts and functions thereof are related at intermediary levels directly to corresponding states of the system itself or of the external world.

From all of this, one can conclude that the order between the parts should in fact rather be the *opposite*. See Figure 2.

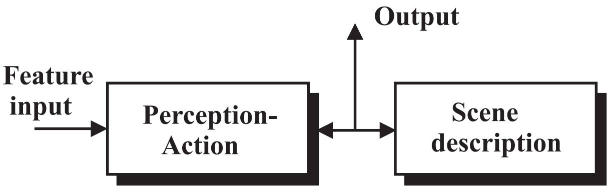

Fig. 2. Perception-Action robotics model

From these considerations, a system structure is proposed, where the first part of the system, step by step performs a mapping from percepts onto actions or states. The central mechanism is the *perception-action* feedback cycle. The usual assumption that certain percepts shall lead to corresponding learned actions is true, for normal operation of a *trained* system. However, in the learning phase, we have the situation that *action precedes perception* for the part of the structure involved in the learning.

The crucially important reason for this inverse causal direction is that action or state space is much less complex than percept space. The number of possible combinations of perceptual primitives in an image is huge, and most combinations of these will not be of interest as they will never occur. It is necessary to identify the combinations which may occur as economically as possible. Given that the state space is less complex, and that feature combinations of interest will be specifically apparent as the system moves around in the state space, this movement of the system in the state space can be used to organize the relevant parts of the feature space, and associate them to the generative states. Although being a traditional view, the opposite could never be possible, due to the tremendous complexity of the percept space.

The system will in the early bootstrapping learning phase be driven by partly random actions, which however are known to the system itself. This will implement elements of exploration, where the system can observe how the percept structure changes, and perform an associative learning of these relations. Starting with a simple environment, it is easy to distinguish which percepts change as a function of an action or a state change. 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 of 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.

Driving a learning system using semirandom signals for organization of the nervous system, is a well known mechanism from biology. Many low level creatures have built in noise generators, which generate muscle twitches at an early stage of development, in order to organize the sensorial inputs of the nervous system. Organization is driven from the motor side and not from the sensing side. It has been convincingly shown that noise and spontaneously generated neural activity is an important component to enable organization and coordinated behavior of organisms [14].

The major issue is that percepts shall be mapped directly onto states or responses or functions involving these, rather than onto higher level model abstractions. We will see that this step-wise relation of the percept structure to the states of the system or the external world, reduces the complexity at each interface between the two domains. This makes an association structure feasible, as the local system complexity becomes limited. The limited number of degrees of freedom allows the implementation of self-organizing or learning procedures, as there is a chance to perform fast converging optimizations. This gives a mechanism to step by step build up what finally may become a very complex processing structure, useful for the handling of a very complex niche of the external world. See also [5, 8]. In contrast, it would never be possible to build up a complex processing structure, with a large number of degrees of freedom in a single learning or optimization phase.

Driving the system using response signals has four important functions:

- *To separate different modalities of an object from each other, such that they can be separately controlled.*
- *To identify only the percepts which are related to a particular action modality*
- *To provide action outputs from the network generated. Without a response output path, it remains an anonymous mode unable to act into the external world.*
- *Related points in the response domain exhibit a much larger continuity, simplicity and closeness than related points in the input domain.*

It is necessary that the network structure generated has an output to generate responses, which may be an activation of other structures outside the network, or to produce actions. If no such associations could be made, the network in question would have no output and consequently no meaning to the structure outside. This relates to the advantages of the *view centered object representation*, which will be discussed in a subsequent section.

3.1 A Continual Learning Process

It should be emphasized that a system with this strategy will not simply switch between two modes, where it is either learning, or in operation using its acquired knowledge. Rather it will simultaneously use both strategies, and in effect learn most of the time — at some level. We have earlier emphasized that learning takes place through exploration, and in that process action precedes perception. This process only takes place in a limited part of the system, at a given instance. In that learning process, the system uses the competences it has acquired from earlier training experiences. We will use the following terminology:

By *selfgenerated action* we denote an activity of a system, which is produced without any apparent external influence. The action is assumed to be caused by a random noise signal affecting a choice or creating an activity in parts of the system.

By *reaction* we mean that the system performs an activity which is initiated or modified by a set of percepts.

It should be noted that a particular action under consideration, may typically be composed by both selfgenerated components and reactive components. This is given an intuitive illustration in Figure 3.

In the training to deal with a certain phenomenon, here indicated at the highest level, the system associates percepts to the states which are changing. In this process, it uses earlier acquired information in a normal mapping of percepts onto actions in the lower levels. However, at the level to be trained, the action states constitute the organizing mechanism for the associations to implement.

Different parts in the structure use either of the two mechanisms at the same time:

- **Exploratory behavior**: Selfgenerated action → Percept → Association
- **Reactive behavior**: Percept → Reaction

A selfgenerated action by the system causes a particular set of percepts to appear. The actions and percepts are linked to each other in an association process.

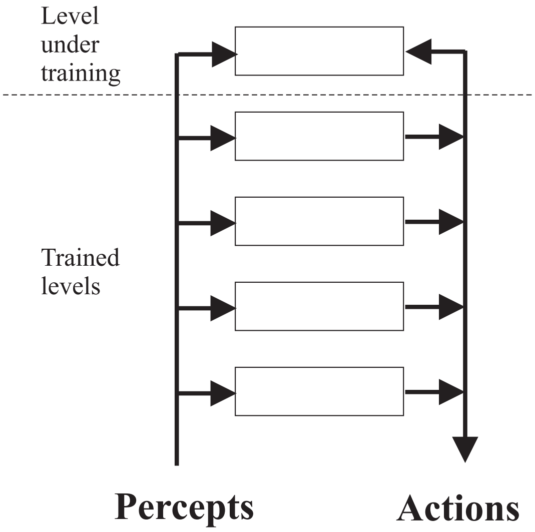

Fig. 3. Intuitive illustration of hierarchical training procedure

This means that at any given time, most processes in a cognitive system implement a purposeful mapping of percepts onto actions according to the models acquired at training. However, there will generally be a (often highest) level of exploration, which is characterized by random action activities. These random components are propagated through the trained levels, where they perform purposefully within the domains of the experience acquired.

There are strong indications that the association of motor signals or actions and percepts, in the human brain, takes place in the posterior parietal cortex [13, 19, 22, 29, 16, 23]. The principle of a combination between a probing into the external space and sensing, has recently received theoretical support [17].

The mixed exploratory/reactive strategy appears to be true even for cognitive systems of the category Humans: There is always a top level which deals with phenomena hitherto unknown to us and can only be subjected to exploration, implemented as pseudo-random action components at that level. These components are propagated down through the entire multi-level perception-to-action machinery acquired from experience, and implemented as valid action sequences for a sophisticated exploration of a new domain of the environment.

3.2 Interface Between Percept-Action and Symbolic Representation

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 [24, 4, 25]. It has also been found that the perception of an object results in the generation of motor signals, irrespective of whether or not there is an intention to act upon the object [1, 10].

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.

This motivates us to propose a more detailed structure for a technical cognitive vision system, as given in Figure 4.

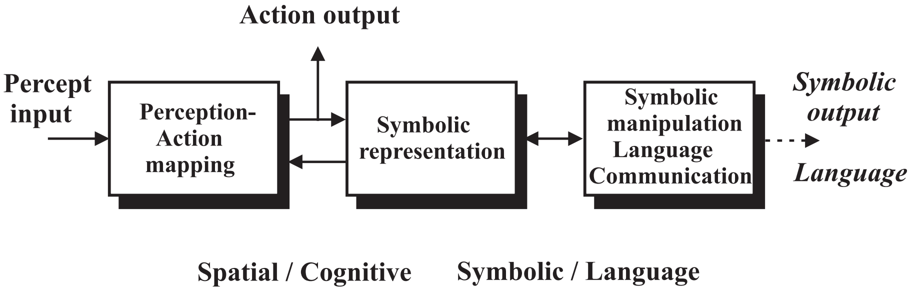

Fig. 4. Perception-action robotics model

The perception-action mapping part is attached to the the part for symbolic processing through an interface, which will be the subject of this section. So far the discussion may have implied that we would have a sharp division between a percept side and a response side in the structure. This is certainly not the case. There will be a mixture of percept and response components to various degrees in the structure. We will for that purpose define the notion of *percept equivalent* and *response equivalent*. A response equivalent signal may emerge from a fairly complex network structure, which itself comprises a combination of percept and response components to various degree. At low levels it may be an actual response muscle actuation signal which matches or complements the low level percept signal. At higher levels, the response complement will not be a simple muscle signal, but a very complex structure, which takes into account several response primitives in a particular sequence, as well as modifying percepts. The designation implies a complementary signal to match the percept signal at various levels. Such a complex response complement, which is in effect equivalent to the system state, is also what we refer to as *context*.

A response complement also has the property that an activation of it may *not necessarily* produce a response at the time, but rather an activation of particular substructures which will be necessary for the continued processing. It is also involved in knowledge acquisition and prediction, where it may not produce any output.

The simple block structure of Figure 4 is obviously not fair to the assumed complexity. A variety is given in Figure 5, which is is intended to illustrate the fact that certain classes of percepts may map onto actions after a very brief processing path, much like what we know as reflexes in biological systems. Such direct mapped actions may however require a fairly complex contextual setup, which determines the particular mapping to use. Other actions may require a very complex processing, involving several levels of abstraction. This general, distributed form of organization is supported by findings in biological visual systems [29, 26].

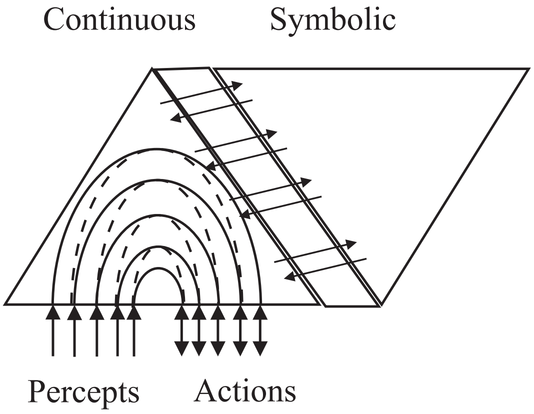

Fig. 5. Pyramid version of the perception-action robotics model

There are other important issues of learning such as representation of purpose, reinforcement learning, distribution of rewards, evolutionary components of learning, etc, which are important and relevant but have to be omitted in this discussion.

The number of levels involved in the generation of a response will depend on the type of stimulus input as well as of the particular input. In a comparison with biological systems, a short reflex arch from input to response may correspond to a skin touch sensor, which will act over interneurons in the spinal cord. A complex visual input may involve processing in several levels of the processing pyramid, equivalent to an involvement of the visual cortex in biological systems.

The assumption stated in the previous section is that the symbolic representation shall be derived from the states or actions generated from the perception-action part. From biological systems it is known that these actions may be in fact “acted out”, or they may be inhibited from generating a physical action. In either case, they represent the interpretation of the set of percepts, which can be used for a symbolic processing [1, 10].

Representation of context at lower levels of a cognitive system is more complex and spatial/quantitative than we are used to for symbolic descriptions. Symbolic 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.

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 as symbols 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.

4 Symbolic Representation and Processing

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, temporal and 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 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 relatively low information content, because it maps onto a rich spatial knowledge structure at all levels, available within our surprisingly similar brain structures. This information, however, derives from the individual exploration of what are similar environments.

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 actual contextual and action parameters are reinserted in a way related to the current 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 current state of a system, be it the own system or another system.

It is postulated that *metric* or similarity measures are only available within the spatial-cognitive part of a cognitive system and not in the symbolic part. This means that as two symbolic entities are to be compared, they are communicated from the symbolic part to the spatial-cognitive part of the system, where the comparison is implemented. One reason for this is that a relevant comparison usually requires the inclusion of actual quantitative parameters.

The output from the symbolic processing is normally (for biological systems) output over the action output interface according to Figure 4. For a technical system, there may be possibilities to interface directly to to the symbolic processing unit, as indicated by the dashed line to the right. The reason is that the symbolic representation used in this part of the system, can be expected to be less context dependent and consequently more invariant than what is the case in the perception-action part. This will make it potentially feasible to link symbol states to an outside representation, using association or learning.

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. This fact is also bringing about a shift from the traditional *declarative* representation towards the use of *procedural* representation, allowing association or learning.

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. In particular, it is the difference between a very context specific world and a more invariant and generalizable world.

4.1 Representation in Language

Although the relation to language is not central in this document, we will make a few observations which extend the issues already dealt with. The major part of our dicourse above, is on what is postulated to happen in the spatial-cognitive or procedural part of a vision system, which for human vision is not assumed to be available to us in conscious experience, except for its effects. What is happening in the motor/language/consciousness or declarative part of the human system, on the other hand, is the generation of a normalized object centered representation, in order to be able to communicate it in a sufficiently compact way. In this case it is necessary for compactness to cut off a major part of incidental contextual links, which are probably not necessary for the receiver, as it will not be in exactly the same state as the sender of the message, anyway. The formalism that we find in classical knowledge-based systems is oriented towards such compact, string-representable phenomena intended for communication. As for all object centered representations, taxonomies are built up with an overview of the final outcome, rather than the type of incremental, "blind" buildup which is assumed necessary for view centered representations.

There is a clear difference between what we represent in language as declarative statements, compared to the procedural statements required for generation of responses. While *subset-of* and *member-of* concepts are important for conscious taxonomy and organization, such as to determine a particular disease from its symptoms, it is not apparent that these concepts are useful for response generation systems. The type of grouping or abstraction which is performed here, is in fact similar to the process of increased abstraction which we will see in an object centered representation in contrast to a view centered representation; a number of specific action references are cut off from the representation structure.

The fact that language can communicate action, is because it is derived from action primitives as discussed earlier. The reason language works is due to the rich structure that it evokes in the receiver's cognitive system; not due to the content of the sentence itself. Language should be viewed as a pushing of buttons on the receiver. Most of the information necessary for the response has to be contained in the structure of the receiver; it can not just point to abstract references thereof. This is the major reason for the limited success of inference systems using natural language in robotics: There is too little information contained in language itself, and if there is no powerful spatial-cognitive interpretation structure available, such that language control can be implemented in a sufficiently flexible way.

5 Object-centered versus View-centered Representation

The earlier discussion of two different domains for processing, the perception-action domain and the symbolic domain, has an important relation to two different approaches to object representation: *view-centered* and *object-centered* representation, respectively [21, 8]. See Figure 6.

From a real object, a number of measurements or projections are generated. They may e.g. be images of the object, taken from different angles. See Figure 6a. From these measurements we can proceed along either one of two different tracks.

One of the tracks leads to the object-centered representation which combines these measurement views into some closed form mathematical object [11]. See Figure 6b. The image appearance of an instance of a particular orientation of the object is then obtained using separate projection mappings.

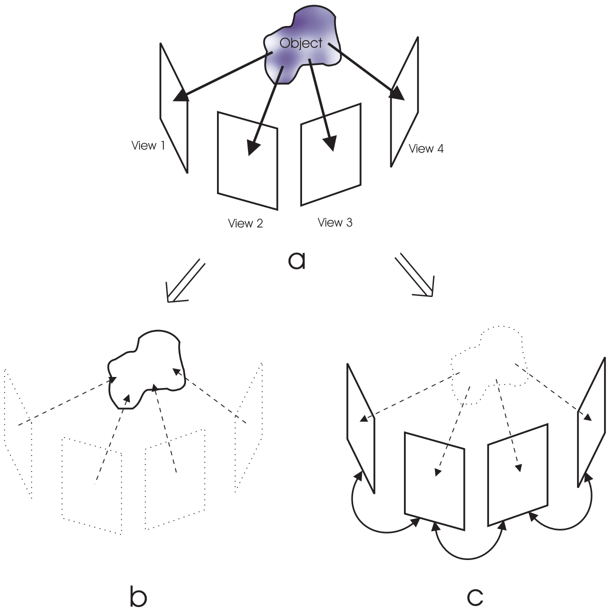

Fig. 6. Object-centered and view-centered representation of an object. a) Measurements produce information about different views or aspects of an object. b) Object-centered representation: The views are used to reconstruct a closed form object representation. c) View-centered representation: The views are retained as entities which linked together form a representation of the object.

A view-centered representation, on the other hand, combines a set of appearances of an object, without trying to make any closed form representation [27, 18, 2]. See Figure 6c.

5.1 Object-Centered Representation

The basic motive of the object-centered representation is to produce a representation which is as compact and as invariant as possible. It generally produces a closed form representation, which can subsequently be subjected to interpretation. This implies that no unnecessary information is included about details on how the information was derived. A central idea is that matching to a reference should be easier as the object description has no viewpoint-dependent properties. A particular view or appearance of the object can be generated using appropriate projection methods.

We can view the compact invariant representation of orientation as vectors and tensors [7], as a simple variety of object-centered representation. Over a window of a data set, a set of filters are applied producing a component vector of a certain dimensionality. The components of the vector tend to be correlated for phenomena of interest, which means that they span a lower dimensional sub-space. The components can consequently be mapped into some mathematical object of a lower dimensionality, to produce a more compact and invariant representation, i.e. a vector or a tensor [7].

A drawback of the object-centered representation is that it requires a preconceived notion about the object to ultimately find; its mathematical and representational structure, and how the observed percepts should be integrated to support the hypothesis of the postulated object. It requires that the expected types of relations are predefined and already existing in the system, and that an external system keeps track of the development of the system such as the allocation of storage, and the labeling of information. Such a preconceived structure is not well suited for self-organization and learning. It requires an external entity which can "observe labels and structure", and take action on this observation. It is a more classical declarative representation, rather than a procedural representation.

5.2 View-Centered Representation

In a view-centered representation, no attempt is made to generalize the representation of the entire object into some closed form. The different parts are kept separate, but linked together using the states or responses, which correspond to or generated the particular views. This gives a representation which is not nearly as compact or invariant. However, it tells what state of the system is associated to a particular percept state. A view-centered representation in addition, has the advantage of being potentially self-organizing. This property will be shown to be crucial for the development of a learning percept-action structure. There are indications from perceptual experiments, that the view-centered representation is the one used in biological visual systems [21].

An important characteristic of the view representation is that it directly implements an *interpretation*, rather than a geometrical *description* of an object that we want to deal with. By interpretation we denote links to actions that are related to the object, and information about how the object transforms under the actions. This is what motivates the choice of structure described in earlier sections.

5.3 Combination of Representation Properties

An object centered representation is by design as invariant as possible with respect to contextual specificities. It has the stated advantage to be independent of the observation angle, distance, etc. This has, however, the consequence that it cuts off all links that it has to specific contexts or response procedures which are related to that context or view. We recognize this from the preceding section as what is characterized as a *symbolic* representation. The generation of an invariant representation, implies discarding information which is essential for the system to act using the information. In order to use such information, a system has to introduce actual contextual information.

It is postulated that an interpreting system shall start out from the view centered representation of objects. This allows the system to represent phenomena such as objects as combinations of percepts and responses. It is furthermore postulated that these can be viewed as *invariants*, i.e. there is a local region in some subspace in which small changes in the percept domain and the action or state domain are equivalent.

The structure which results from the preceding model will be of type frames-within-frames, where individual transformations of separate objects are necessary within a larger scene frame. The ability to handle local transformations is absolutely necessary, and would not be possible with a truly iconic view representation. It is postulated that the frames-within-frames partitioning is isomorphic with the response map structure. In this way, the response map “reaches” into the frame in question, to implement the percept-action invariance of a particular object aspect.

6 Extending and Exchanging Competences of Vision Systems

In the preceding sections, most of the attention has been paid to exploratory learning. While this modality is basic, there are other modalities available for the extension of the competences of a system. A system can acquire information using three different mechanisms, where each one is dependent upon the acquisition in the previous mechanisms:

1. Copying at the time of system generation
2. Active exploration, defining concept spaces through associative learning
3. Passive observation or communication, i.e. using language, can be used to link points in available concept spaces

6.1 Copying at the time of system generation

Any system will start with some set of basic hardware and software in some representation. It is in principle possible to copy an existing system to generate another system identical to the first one. There may well in a given case be practical complications, which may make such a procedure prohibitive. The reason why it is in principle possible, is that a copying procedure requires no *interpretation* of the structure it deals with, only a “blind” but accurate copying mechanism. A device of a similar character is a TV set, which maps the points of an image from one surface to another, not ever worrying about the content of the image, what parts relate to each other, etc.

In contrast, the task to copy *a certain item* of information from one system to another is generally not possible. Transfer of an item of information from one system to another must be made through the normal in- and outputs of the systems, and on the terms of the systems. There are two reasons for this:

1. As soon as a system takes on to explore the external world, it will proceed along a different trajectory and acquire different information than another system, even if they started out identical. Systems with the architecture proposed will store information in a *semantic* form, which means in relation or adjacency to similar information. This means that the sequence of exposure to different objects in the environment will lead to different organizations of the information. It is consequently not possible to map the information from one system to another, even if they initially were identical, which means that it is not obvious where the information should be "pushed into" the receiving system.
2. It is not clear how to identify the information belonging to a certain item or topic, as there will be attachments to references and context. All these links would be required to be identified and labeled in order to be possible to reinsert in the receiving system. It is not clear how such an establishment of correspondence could ever be made in a self-organizing knowledge structure, be it in the receiving system or in the donating system. The receiving system may in fact lack some of the contextual parameters required for attachment.

We can conclude that that the only mode for transmission of information from one system to another is over the normal input and output channels, on the terms of the receiving system. The last qualification implies that the just received information has to be stored in relation to similar, earlier acquired information and context. This has the consequence that the possibility for a designer to "push in" information into an operating system, is less than traditionally desired. Such information has to go through the interpreting and organizing mechanisms of the system, as discussed above.

The way that the "practical complications" of copying have been resolved for biological systems, is that a "blueprint" of the system is copied, rather than the system itself. This has the consequence that very limited quantities of the experience of a system can be transferred.

6.2 Active exploration, defining concept spaces

Most of what has been discussed in earlier sections deals with the acquisition of information through active exploration. This is the fundamental mode of learning, as it establishes the basic relations between action or state and available percepts. It is postulated that this explorative perception-action mapping defines *concept spaces*, defining domains of phenomena it can deal with. The characteristics of an object are consequently the way its appearance behaves under different modes of responses. These are the different percept-response invariants referred to in [5]. The crucial issue is that these invariants are determined by the response modes of the observing system. The associative learning system will create these invariants to "match" the response effects. As a consequence, the ability of the system to understand the world is given by its ability to manipulate it. The preceding leads to the basic principle:

> *A system's ability to interpret objects and the external world is dependent upon its ability to flexibly interact with it.*

A philosophical consequence of this is that for a reasonable degree of intelligence to develop in a system, it requires a flexible machinery for a complex interaction with the external world, in addition to an effective information processing architecture. Another consequence of the necessity to disturb an object, in order to understand it, leads to a variety of the Uncertainty relation [28].

6.3 Passive observation or communication

An apparent question is now: In order for the system to be able to deal with objects, is it necessary for it to interact with every such single object in every respect?

It is postulated that a response driven learning session will define sets of concept spaces. The system can then be expected to deal with objects or cases which are *sufficiently similar* to what it has experienced in the response driven learning process, through interpolation and extrapolation within the defined spaces. Exactly what sufficiently similar implies is not clear at this point. Abstractly, it must imply that an earlier defined percept-response space is valid for the phenomenon under consideration, although it may be parametrically different from what the learning trajectory defined. In some sense, it must imply that the problem structure or topology is similar. This is similar to Piaget's concepts of *assimilation* and *accommodation* in his theory of cognitive development.

It is well supported that humans are subject to the same limitations and possibilities. We can comprehend a phenomenon which we have not experienced before, as long as it contains components which are sufficiently similar to something of which we have an interactive experience; i.e we can deal with it as an interpolation or extrapolation within an already available percept-response concept space. There are indications that at an adult age, most of the concept spaces used are already available, and that most of our knowledge acquisition after this deals with combinations of particular instances, and interpolations or extrapolations within these available spaces. This can conceivably imply that cases with the same problem structure or within the same concept space, may well appear very different, but we can handle them without considerable difficulty.

In this way, a system can comprehend a phenomenon from passively observed imagery, as long as the primitive components of it can be mapped over the association concept spaces available to the system. Similarly, language can evoke components which are familiar to the system, although the particular arrangement of them may be new.

In such a way, knowledge may be communicated in an efficient way to a system, such that it does not need to make an involved experience, but can observe passively or communicate symbolically i.e. using some language.

7 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, as it implies an embodiment which can interact with the external world. Does that mean that the preceding cognitive vision architecture is only applicable to robotics?

Not exclusively! The belief is however that the structure discussed is not only advantageous, but inevitable, for the development of demanding applications of vision. This also includes the interpretation of complex static imagery.

It will similarly be inevitable for cases where the output is not a physical action but the communication of a message. An example of the latter type is complex 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 systems 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.

From this derives the view that the development of powerful technical vision 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. 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 in cognitive vision.

8 Implementation of Cognitive Mechanisms in Vision Systems

The principles discussed in this paper have been tested in a structure for unrestricted recognition of 3-D objects, documented separately in [9]. By unrestricted, we imply that the recognition shall be done independently of object position, scale, orientation and pose, against a structured background. It shall not assume any preceding segmentation and allow a reasonable degree of occlusion.

The paper [9] describes how objects can be represented. A structure is proposed to deal with object and contextual properties in a transparent manner. The method uses a hierarchy of triplet feature invariants, which are at each level defined by a learning procedure. In the feed-back learning procedure, percepts are mapped upon system states corresponding to manipulation parameters of the object. The method uses a learning architecture employing channel information representation [6].

9 Concluding Remarks

There is a traditional belief that percepts are in some way "understood" in a vision system, after which suitable responses are attached. This does however require simple units to have an ability of "understanding", which is not a reasonable demand upon local structures. This is a consequence of the luxury of our own capability of consciousness and verbal logical thinking. This leads to a potential danger inherent in our own process for system development, in that our conscious experience of the environment has its locus at the very end of the chain of neural processing; at a point where in fact most of the cognitive processing has already been made. It is well established that language capability, logical reasoning and conscious processing are all derivates of the motor system normally in the left half of the brain, at the end of the processing chain.

A processing in terms of such conscious object terms *inside* the cognitive processing structure is not likely to occur. Most, or nearly all of the cognitive processing has already been done when signals are available for the motor manipulation objects to take place, or a representation in conscious terms to emerge. We are too late to see *how* it happened. We only notice *what* has happened.

Given the well known distributed nature of processing, it is apparent that there is no basis for any estimation of importance or "meaning" of percepts locally in a network, but that "blind and functional rules" have to be at work to produce what is a synergic, effective mechanism. One of these basic rules is undoubtedly to register how percepts are associated with responses, and the consequences of these. This seems at first like a very limited repertoire, which could not possibly give the rich behavior necessary for intelligent systems. Like in the case of other evolutionary processes for self-organization, there seems to be no other credible alternative. Rather, we have to look for simple and robust rules, which can be compounded into sufficient complexity to deal with complex problems in a "blind" but effective way.

Acknowledgments

The author wants to acknowledge the financial support of WITAS: The Wallenberg Laboratory For Information Technology and Autonomous Systems, SSF (The Foundation for Strategic Research) as well as the Swedish National Board of Technical Development. Considerable credit should be given to the staff of the Computer Vision Laboratory of Linkoeping University, for discussion of the contents of this paper.

References

1. S. J. Anderson, N. Yamagishi, and V. Karavia. Attentional Processes link perception and action. *Proc R Soc Lond B Biol Sci.*, 269(1497):1225–1232, 2002.
2. D. Beymer and T. Poggio. Image Representations for Visual Learning. *Science*, 272:1905–1909, June 1996.
3. Antonio Chella, Marcello Frixione, and Salvatore Gaglio. A cognitive architecture for artificial vision. *Artificial Intelligence*, 89(1–2):73–111, 1997.
4. M. S. Gazzaniga. *The Social Brain. Discovering the Networks of the Mind.* Basic Books, New York, 1985.
5. G. H. Granlund. The complexity of vision. *Signal Processing*, 74(1):101–126, April 1999. Invited paper.
6. G. H. Granlund. An Associative Perception-Action Structure Using a Localized Space Variant Information Representation. In *Proceedings of Algebraic Frames for the Perception-Action Cycle (AFPAC)*, Kiel, Germany, September 2000.
7. G. H. Granlund and H. Knutsson. *Signal Processing for Computer Vision.* Kluwer Academic Publishers, 1995. ISBN 0-7923-9530-1.
8. Gösta Granlund. Does Vision Inevitably Have to be Active? In *Proceedings of the 11th Scandinavian Conference on Image Analysis*, Kangerlussuaq, Greenland, June 7–11 1999. SCIA. Also as Technical Report LiTH-ISY-R-2247.
9. Gösta H. Granlund and Anders Moe. Unrestricted Recognition of 3-D Objects for Robotics Using Multi-Level Triplet Invariants. *Artificial Intelligence Magazine*, 2003. To appear.
10. J. Grezes, M. Tucker, J. Armony, R. Ellis, and R. E. Passingham. Objects automatically potentiate action: an fmri study of implicit processing. *Eur J Neurosci*, 17(12):2735–2735, 2003.
11. W. E. L. Grimson. *Object Recognition by Computer: The Role of Geometric Constraints.* MIT Press, Cambridge, MA. USA, 1990.
12. R. M. Haralick and L. G. Shapiro. *Computer and Robot Vision.* Addison-Wesley, 1992.
13. Marc Jeannerod. *The Cognitive Neuroscience of Action.* Blackwell publishers Ltd, 1997.
14. L. C. Katz and C. J. Shatz. Synaptic Activity and the Construction of Cortical Circuits. *Science*, 274:1133–1138, November 15 1996.
15. Tai Sing Lee and David Mumford. Hierarchical Bayesian inference in the visual cortex. *Optical Society of America*, 20(7):1434–1448, July 2003.
16. Kenji Matsumoto, Wataru Suzuki, and Keiji Tanaka. Neuronal Correlates of Goal-Based Motor Selection in the Prefrontal Cortex. *Science*, 301(5630):229–232, 2003.

17. D. Philipona, J.K. O'Regan, and J.-P. Nadal. Is There Something Out There? Inferring Space from Sensorimotor Dependencies. *Neural Comp.*, 15(9):2029–2049, 2003.
18. T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.
19. Javier Quintana and Joaquin M. Fuster. From Perception to Action: Temporal Integrative Functions of Prefrontal and Parietal Neurons. *Cereb. Cortex*, 9(3):213–221, 1999.
20. Rajesh P. N. Rao and Dana H. Ballard. An active vision architecture based on iconic representations. Technical Report TR548, 1995.
21. M. Riesenhuber and T. Poggio. Computational models of object recognition in cortex: A review. Technical Report 1695, Artificial Intelligence Laboratory and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Aug. 2000.
22. E.M. Robertson, J.M. Tormos, F. Maeda, and A. Pascual-Leone. The Role of the Dorsolateral Prefrontal Cortex during Sequence Learning is Specific for Spatial Information. *Cereb. Cortex*, 11(7):628–635, 2001.
23. Andrew B. Schwartz, Daniel W. Moran, and G. Anthony Reina. Differential Representation of Perception and Action in the Frontal Cortex. *Science*, 303(5656):380–383, 2004.
24. R. W. Sperry. *Science and Moral Priority: Merging Mind, Brain and Human Values*. Praeger, New York, 1985.
25. S. P. Springer and G. Deutsch. *Left Brain, Right Brain*. Freeman, New York, 1993.
26. German Sumbre, Yoram Gutfreund, Graziano Fiorito, Tamar Flash, and Binyamin Hochner. Control of Octopus Arm Extension by a Peripheral Motor Program. *Science*, 293(5536):1845–1848, 2001.
27. S. Ullman and R. Basri. Recognition by linear combinations of models. 13(10):992–1006, 1991.
28. R. Wilson and G. H. Granlund. The Uncertainty Principle in Image Processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI–6(6), November 1984. Report LiTH-ISY-I-0576, Computer Vision Laboratory, Linköping University, Sweden, 1983.
29. Ulf Ziemann. Sensory-motor integration in human motor cortex at the premotoneurone level: beyond the age of simple MEP measurements. *J Physiol (Lond)*, 534(3):625–, 2001.