From Multidimensional Signals to the Generation of Responses

Gösta H Granlund

Computer Vision Laboratory, Department of Electrical Engineering,
Linköping University, S-581 83 Linköping, Sweden

Abstract. It has become increasingly apparent that perception cannot be treated in isolation from the response generation, firstly because a very high degree of integration is required between different levels of percepts and corresponding response primitives. Secondly, it turns out that the response to be produced at a given instance is as much dependent upon the state of the system, as the percepts impinging upon the system. The state of the system is in consequence the combination of the responses produced and the percepts associated with these responses. Thirdly, it has become apparent that many classical aspects of perception, such as geometry, probably do not belong to the percept domain of a Vision system, but to the response domain.

There are not yet solutions available to all of these problems. In consequence, this overview will focus on what are considered crucial problems for the future, rather than on the solutions available today. It will discuss hierarchical architectures for combination of percept and response primitives, and the concept of combined percept-response invariances as important structural elements for Vision. It will be maintained that learning is essential to obtain the necessary flexibility and adaptivity. In consequence, it will be argued that invariances for the purpose of vision are not geometrical but derived from the percept-response interaction with the environment. The issue of information representation becomes extremely important in distributed structures of the types foreseen, where uncertainty of information has to be stated for update of models and associated data.

1 Introduction

A fundamental problem is how to assemble sufficiently complex models and the computational structures required to support them. In order for a system modeling a high structural complexity to be manageable and extendable, it is necessary that it exhibits modularity in various respects. This implies, for example, standardized information representations for interaction between operator modules. One way to satisfy these requirements is to implement the model structure in a regular fashion as a *hierarchy*, although we should bear in mind that the communication need not be restricted to adjacent layers of such a hierarchy.

We can distinguish between two different types of hierarchies:

- *Scale hierarchies*
- *Abstraction hierarchies*

Most of the work on pyramids so far has dealt with size or scale, although they have indirectly given structural properties. They will not be dealt with in this paper, but descriptions can be found in [6,15,21].

Granlund introduced an explicit abstraction hierarchy [4], employing symmetry properties implemented by Gaussian envelope functions, in what today is commonly referred to as Gabor functions or wavelets [2]. An *abstraction hierarchy* implies that the image can be considered as an expansion into *image primitives*, which can be viewed as conceptual building blocks forming the image. In this concept lies the assumption of a hierarchy, such that building blocks at a lower level form groups which constitute a single building block at a higher level. Building blocks at the two levels are viewed as having different *levels of abstraction*.

Fig. 1. Conceptual representation of an image as an abstraction hierarchy.

As an example we can look at Figure 1, which suggests a particular set of abstraction levels. At the lowest level we assume the image itself, describing a distribution of density and possibly color. At the second level we have a description of line and edge elements. At the third level we may have a description of curvature, or convexity and concavity. At the fourth level we may have outlines. At the fifth level we may have relations between objects, and continue to higher levels as appropriate.

It is necessary to omit the important discussion of mathematical representations for information in this document. It turns out that for 2-D information a *vector* representation is advantageous, while for 3 dimensions and higher, *tensors* will do the work. These representations allow the use of certainty statements for all features, which can be updated with respect to models and data. For further details, reference has to be made to [6].

In the same way that context information affects our interpretation of a more local event, information at a higher level can be used to control the processing at a lower level. This is often referred to as *top-down* processing. A hierarchical structure allowing this is illustrated intuitively in Figure 2.

Fig. 2. A hierarchical structure with bottom-up and top-down flow of information.

In such a structure, processing proceeds in a number of stages, and the processing in one stage is dependent upon the derived results at higher levels. This leads to a model based analysis, where models are assembled from combinations of primitives from several levels. An important property is that these models do not remain constant, but adapt to the data, which can be used for adaptive fil-

tering in multiple dimensions [8]. This is a very important issue which, however, goes beyond the objectives of this document, and reference has to be made to [6].

2 Similarity Representations for Linked Structures

Most information representation in vision today is in the form of arrays. This is advantageous and easily manageable for stereotypical situations of images having the same resolution, size, and other typical properties equivalent. Increasingly, various demands upon flexibility and performance are appearing, which makes the use of array representation less obvious.

The use of actively controlled and multiple sensors requires a more flexible processing and representation structure, compared to conventional methodology. The data which arrives from the sensor(s) can be viewed as patches of different sizes, rather than frame data in a regular stream and a constant array arrangement. These patches may cover different parts of the scene at various resolutions. Some such patches may in fact be image sequence volumes, at a suitable time sampling from a particular region of the scene, to allow estimation of the motion of objects. The information from all such various types of patches has to be combined in some suitable form.

The conventional array form of image information is in general impractical as it has to be searched and processed when some action is to be performed. It would be desirable to have the information in some suitable, partly interpreted form to fulfill its purpose to rapidly evoke actions. The fact that the information has to be in some interpreted form, implies that it should be represented in terms of content or *semantic* information, rather than in terms of array values. As we will see, content and semantics implies relations between units of information or symbols. It is consequently attractive to represent the information as *linked objects*. The discussion of methods for representation of objects as linked structures will be the subject of most of this document, and we can already now observe how some important properties of such a representation relate to that of conventional array representations:

- The array implies a given size frame, which can not easily be extended to incorporate a partially overlapping frame
- Features of interest may be very sparse over an array, leaving a large number of unused positions in the array

There is a great deal of literature available on the topic of object representation using classical methods[7], which however will not be reviewed here. Most of these methods treat objects with respect to geometric properties expressed in some coordinate system. They relate to rules of interpretation, which should be input into the system. This is probably appropriate for the representation of low level properties. For higher level properties the requirements are different.

2.1 Continuity, Similarity and Semantics

In the world around us, things generally appear different, whether they are or not. A particular object will appear different seen from different angles. Still we can recognize most objects at arbitrary positions, orientations, distances, etc. An object which persists in appearing different from anything else we know, can not be matched to any known class, which is a common purpose of recognition. There have to be aspects which are sufficiently familiar, for us to start a process of recognition. For that reason we will be interested in the *simultaneous appearance of similarity and difference* of properties of objects. This is related to notions which will be termed *invariance* and *equivariance* for future discussions.

Along the lines spelled out earlier, we can view the detection of invariance as the result of a test establishing similarity, while equivariance is the result of a test establishing a distance measure.

The representation of information in a cognitive system is crucial for effective performance. Traditionally, the representation is in the form of natural language, which has the following less desirable properties:

- Discrete and discontinuous: Similarity is established by matching, and the result is **MATCH** or **NO MATCH**
- Non-metric: It is not possible to establish a degree of similarity or distance between symbols

As an example we can take the words:

Fig. 3. Example of words having a small distance in terms of an ASCII letter metric, but large distances in content or meaning

Establishing a similarity measure e.g. using their ASCII numerical value would not be useful. Such a representation can not be used for efficient processing of semantic or cognitive information.

We can conclude that a suitable representation for semantic information requires:

Continuous representation of similarity in content

In the preceding case with words, we can observe that we deal with two types of adjacency:

- Time or position adjacency between words
- Content adjacency or similarity in meaning between words

It is apparent that both of these distance measures have to be represented in the description, although this is not the case in the example above. It is fairly obvious what the space or time distance represents, but what about the similarity in property? We will address this question in the next section.

2.2 Channel Representation of Information

A representation of similarity assumes that we have a distance measure between items. For an advanced implementation of a linkage structure, we assume that information is expressed in terms of a *channel representation*. See Figure 4.

Fig. 4. Channel representation of some property as a function of match between filter and input pattern domain

Each channel represents a particular property measured at a particular position of the input space. We can as a first approximation view such a channel as the output from some band pass filter sensor for some property. If we view the channel output as derived from a band pass filter, we can first of all establish a measure of *distance* or *similarity* in terms of the parameters of this filter. See Figure 4. For a conventional, linear simple band pass filter, the phase distance between the flanks is a constant $\pi/2$. Different filters will have different band widths, but we can view this as a standard unit of similarity or distance, which is supplied by a particular channel filter.

This resembles the function of biological neural feature channels. There are in biological vision several examples available for such properties; edge and line detectors, orientation detectors, etc. Such a channel filter has the characteristic that it is local in some input space, as well as local in some property space.

It may map from some small region in the visual field, and indicate, say, the existence of a line at some orientation.

We can view every channel as an originally independent fragment of some space. It should now be observed that we have two different types of distances:

- *Distance in input space*
- *Distance in property space*

The distance in input space may be the distance between two different positions of a line within the receptive field of an orientation detector, where the line has a constant orientation.

The distance in property space may be the difference between two different orientations of a line, located centrally within the receptive field of an orientation detector.

A variation of position or a variation in orientation will both of them give a variation of the output according to Figure 4, and a priori, we cannot distinguish between these two situations, having a single and simple stimulus acting upon a single orientation detector. Either a line at the proper orientation can be moving over the spatial region of the detector, or a line at the proper position can be rotating over the detector, or most likely a combination of both.

Fig. 5. Visualization of channels in input space as well as property space

This leads us to consider the input space and the property space as two orthogonal subspaces, which in the general case both will contribute to the output in some linear combination. See Figure 5. The distance represented by the

channel filter will be in a linear combination from both of these spaces. Distance is a property which is well defined in a multidimensional space. Distance does not allow us to order events, but to define a sequence of events, represented by nodes which are joined by links. Every such link will represent a different one-dimensional projection from the multidimensional space under consideration, than a joining link.

The fact that we can view the phase distance between two adjacent channel peaks as $\pi/2$, implies that we can view the two subspaces as orthogonal in the metric defined by the joining band pass filter. Still these subspaces are parts of some larger common vector space.

This means that the subspaces which relate to each filter output are different, and cannot really be compared in the same two-dimensional projection plane, as suggested in Figure 5. Each subspace can for intuitive visualization be represented as a vector, which is orthogonal to its nearest neighbor subspaces. This is illustrated in Figure 6. As can be seen from Figure 6, the vectors are tapering off from the center. This indicates that while adjacent subspaces are orthogonal, we cannot say much about the relation between vector subspaces at a larger distance. What we can assume is that the subspaces "bend" into other parts of the common vector space, which makes them disappear from the horizon of any given vector subspace. This can be viewed as a curvature of the local space around a particular subspace, or as a windowing effect. As such, it may well be a necessary introduction of locality providing a stabilizing effect for the resulting networks, much like lateral inhibition.

Fig. 6. Representation of channels as orthogonal subspaces

2.3 Implications of Multiple Measurements

From the previous section it follows that similarity is measured and valid along a single, one-dimensional subspace only, given the output from one single channel. For a particular object, there will be different distance measures to another particular object, in terms of different properties. In addition, any two successive links may not necessarily represent distances along the same, one-dimensional

subspace, as we have no way to track what filters are involved where. There is consequently no way to order objects unambiguously for two different reasons:

1. There is no way to order items which are defined in a multi-dimensional space, which is the *Curse of Multi-dimensionality* [6]
2. It is not possible to sort objects with respect to a particular property, as a similarity between subspaces of different filters can never be established

The potential possibility to sort objects with respect to similarity, to produce a list is consequently not available. The fact that we have different measures of distance between two objects implies that we can represent the objects as points in a sufficiently high dimensional, common space. See Figure 7.

Fig. 7. Distance between two objects measured with two different probes, implying projections upon two different subspaces

2.4 Representation Using Canonical Points

It is postulated that we do not observe the world continuously although it may appear so to us. Rather observations and representations are made in particular, discrete points. We call these *canonical points*.

It is postulated that canonical points relate to certain phases of the features output from the filters involved. It is postulated that canonical points correspond to phases 0, 180 and ± 90 degrees in outputs from these filters. Parenthetically, these values do as well correspond to the discrete eigensystems which are derived from observation operators used for continuous fields in quantum mechanics [10].

It is furthermore postulated that a representation at these characteristic phases gives us exactly the sampling resolution required to provide a sufficiently good description. This can be viewed as a variable sampling density controlled by the content of the image.

It is obvious that there has to be some discretization in the representation of objects and events, implying a certain limited resolution. What is stated here is that this resolution is directly related to the forms and scales of objects and events themselves, mediated by the measurement probes or filters involved. These canonical points will imply different things dependent upon the level and phenomenon under consideration, but in general be points of symmetry, etc. of objects. Canonical points represent what we will abstractly denote an *object*, which in everyday language can be a feature, an edge, a line, an object, an event, a position, a view, a sequence, etc. Every feature or object is provided at some level of the processing hierarchy by something equivalent to a filter. The implementation of this is apparent for low-level features, but we can find equivalent interpretations at higher levels.

3 Representation and Learning of Object Properties

Traditionally it has been assumed that knowledge could actively be input into systems by an operator through specification of rules and instances, an assumption underlying the classical AI approach. It has become increasingly apparent, however, that knowledge cannot be represented as prespecified, designated linkages between nodes, as suggested in common AI symbolic representation. The difficulties of the classical symbolic representation are that:

1. It requires that the available 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.
2. It requires an entity which can “observe labels and structure”, and take action on this observation.

These external, centralized functions make it impossible to have the system itself organize its information.

3.1 Object Invariants Formed by Percept-Response Combinations

Over the years there has been an increasing interest in research on invariants, [22,11,12,14]. Most of the methods proposed treat invariants as geometric properties, the rules for which should be input into the system. Theoretical investigation of invariance mechanisms is certainly an important task, as it will give clues to possibilities and limitations. It is not likely, however, that more advanced invariants can be programmed into a system. The implementation of such invariance mechanisms in systems will have to be made through learning.

Fig. 8. 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

An important application of invariant representation is for object description. There are traditionally two major lines of approach which have been used for object description: *object-centered* and *view-centered* representation. See Figure 8.

An object-centered representation employs the information from a number of views to produce a composite geometrical object [7]. See Figure 8b. The image appearance of an object is then obtained using separate projection mappings. An important idea is that matching can be done more easily as the object description is independent of any viewpoint-dependent properties. A view-centered representation, on the other hand, combines a set of appearances of an object, without trying to make any closed form representation [27,24,1]. See Figure 8c.

We will not make any exhaustive review of the properties of these two approaches, or compare their relative advantages here, but only give some motivation for our choice of what is closer to the view-centered representation. A drawback of the object-centered representation for our purpose 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. Such a preconceived structure is not well suited for selforganization and learning. It is also a more classical geometrical representation, rather than a response domain related representation.

A view-linked representation on the other hand, has the advantage of potentially being self-organizable. There are also indications from perceptual experiments, which support the view-centered representation.

Continuing the discussion of the preceding section, it is postulated that we can advantageously represent objects as invariant combinations of percepts and responses. We will start out from the view centered representation of objects, and interpret this in the light of invariant combinations of percepts and responses.

In order for an entity to have some compact representation, as well as to be learned, it has to exhibit invariance. This means that there has to exist some representation which is independent of the frame in which it is described. The representation must not depend on the different ways it can appear to us. As discussed in the last section, the variation in appearance of views has to be directly related to responses we can make with respect to it.

There are different ways to interpret the combination of views to form an object:

$$\begin{array}{rcl} \textit{View} + \textit{Change in Position} & = & \textit{Invariant} \\ \textit{View} + \textit{View Linkage} & = & \textit{Invariant} \\ \textit{View} + \textit{View Linkage} & = & \textit{Object} \end{array}$$

The combination of views with information concerning the position of these views, which is equivalent to the combination of percepts and responses, should produce an entity which allows an interpretation independently of the angle of observation. This is again equivalent to our notion of an object, as something which is not itself affected by the angle from which we view it.

As an example at a higher level, we can take a robot navigating in a room. The combination of detected corners and objects in the room, and the motion responses which are linking these corners together, constitutes an invariant representation of the room. The fact that a combination is an invariant, will make it interesting as a data object to carry on for further computations.

It is furthermore postulated that the invariance mechanism for the representation of an object as a combination of views and the responses involved, implies a form of *equivalence between structures* in the feature domain and in the response domain. We may say that for the domain of an object, we have a “balance”, or an equivalence between a particular set of features and a particular response.

To emphasize, they are equivalent precisely because the combination of them forms an invariant; an entity whose variation is not perceivable in the combined percept-response domain interface surrounding this description. An invariant inevitably implies the balanced combination of a percept and a response. Thus a given response in a particular system state is equivalent or complementary to a particular percept.

Unless we have to postulate some external organizing being, the preceding must be true for all interfaces between levels where invariants are formed, which for generality must be for all levels of a system. This must then be true for the interface of the entire system to its environment as well. What this implies is that the combination of percepts that a system experiences and the responses it performs, constitute an invariant viewed from the wider percept-response domain. This in turn implies that the entire system appears as an invariant to the environment in this wider domain. To avoid misunderstanding, it has to be emphasized that a system can only be observed externally from its responses, and the effects in this subdomain are as expected not invariant, otherwise the system could not affect its environment.

3.2 Response as the Organizing Mechanism for Percepts

A vision system receives a continuous barrage of input signals. It is clear that the system cannot attempt to relate every signal to every other signal. What properties make it possible to select a suitable subset for inclusion to an effective linkage structure? We can find two major criteria:

1. Inputs must be sufficiently close in the input space where they originate, the property space where they are mapped and/or in time-space. This is both an abstract and a practical computational requirement: It is not feasible to relate events over too large a distance of the space considered. This puts a requirement upon the maps of features available, namely the requirement of *locality*.
2. A response or response equivalent signal has to be available, for three different reasons:
 - The first reason is to provide an indication of motive; to ascertain that there are responses which are associated to this percept in the process of learning.
 - The second reason is to provide a limitation to the number of links which have to be established.
 - The third reason is to provide an output path to establish the existence of this percept structure. Without a response output path, it remains an anonymous mode unable to act into the external world.

From the preceding we postulate that:

The function of a response or a response aggregate within an equivalence class is to produce a set of inputs on its sensors, which similarly can be assumed to belong to a common equivalence class, and consequently can be linked.

In consequence we propose an even more important postulate:

Related points in the response domain exhibit a much larger continuity, simplicity and closeness than related points in the input domain. For that reason, the organisation process has to be driven by the response domain signals.

Signal structure and complexity is considerably simpler in the response domain than in the percept domain, and this fact can be used as a focusing entity on the linkage process, where the system's own responses act as organizing signals for the processing of the input. There is a classical experiment by Held and Hein, which elegantly supports this model [9]. In the experiment, two newborn kittens are placed in each of two baskets, which are hanging in a "carousel" apparatus, such that they are tied together to couple the movements of the kittens. One of the kittens can reach the floor with its legs, and move the assembly, while the other one does not reach the floor and is passively towed along. After some period of time, the kitten which can control its movements develops normal sensory-motor coordination, while the kitten which is passively following the movements fails to do so until being freed for several days. The actively moving animal experiences changing visual stimuli as a result of its own movements. The passive animal experiences the same stimulation, but this is not the result of self-generated movements.

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. There is a traditional belief that percepts are in some way "understood" in a system, after which suitable responses are devised. This does however require simple units to have an ability of "understanding", which is not a reasonable demand upon structures. This is a consequence of the luxury of our own capability of consciousness and verbal logical thinking; something which is not available in systems we are trying to devise and in fact a capability which may lead us astray in our search for fundamental principles. 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.

Driving the system using response signals has two important functions:

- *To simplify, learn and organize the knowledge about the external world in the form of a linked network*
- *To provide action outputs from the network generated*

It is necessary that the network structure generated has an output to allow activation of other structures outside the network. This output is implemented by the linkage to response signals, which are associated with the emergence of

the invariance class. If no such association were made, the network in question would have no output and consequently no meaning to the structure outside.

Driving a learning system using response signals for organization, is a well known function 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. More generally, it is believed that noise is an important component to extend organization and behavior of organisms [13].

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 [16–19].

3.3 Object Properties – Part Percept – Part Response

In the tradition developed within the Vision Community, vision has been the art of combining percepts in a way that will describe the external world as well as possible for purposes of interacting with it. There has been an increasing awareness, however, that perception cannot be treated solely as a combination of perceptual attributes, in isolation from the response generation. As an example, it appears that many classical aspects of perception, such as geometry, most likely do not exclusively belong to the percept domain of a Vision system, but include the response domain. This is supported by recent research about the motor system, and in particular the cerebellum [25].

Invariance mechanisms are central in the description of properties for recognition and analysis. It can be seen as an axiom or a truism that only properties which are sufficiently invariant will be useful for learning and as contributions to a consistent behavior.

To start with, I would like to postulate that the following properties are in fact response domain features, or features dominated by their origin in the response domain:

- Depth
- Geometric transformations
- Motion
- Time

In the percept domain, a plane is not a distinguishable entity. We may perceive the texture of the plane, or we may perceive the lines which limit the plane, and these may be clues, but they do not represent the system's model of a plane. A learning system trying to acquire the *concept of plane*, has to associate perceptual and contextual attributes with a translational movement of the response actuator. This response movement can be a translation laterally along the plane, or it can be a movement in depth to reach the plane. This appears to be the invariance property of a plane, and this invariance property is not located in the percept domain but in the response domain. Similarly, the depth to that plane

or any other structure will correspond to the movements required to reach it. For modeling and prediction, actual movements do not have to be effected, but the equivalent implementation signals can be shunted back into the structure, in a way discussed in the section on representation of time.

Similarly, it is believed that projective transformations are to a large extent response domain features. The reason is that they describe how the external world changes its appearance as a function of our movements in it. The primary step in that modeling is to relate the transformations to ego-motion. A secondary step is for the system to generalize and relate the transformations to a relative motion, be it induced by the system itself or any other cause. This is an important example of equivalence, but also an example of invariance. The system can learn the laws of geometrical transformation as a function of its own responses, and then generalize it to any situation of relative motion of the object in question.

An analogous analysis can be made with respect to motion, and the learning of this representation. Motion is the response domain representation of the material provided by various visual clues.

In the same way, the representation of *time* is postulated to be residing on the response side of the structure as well. What this means is that time is represented in terms of elements of responses. This makes it possible to give a network representation of the phenomenon of time, a representation which does not involve unit delays of some form, but is related to the scale of the time phenomena in a particular context. The network linkages allow us to represent time units in an abstracted way for predictive simulation as well, without actually implementing the responses. A further discussion of the representation of time will be given in Section 4.1.

3.4 Response and Geometry

In the proposed system response and geometry are equivalent. Responses always imply a change of geometry, and geometry is always implemented as the consequence of responses or something directly associated with a response. An associated change of geometry can be implemented as a response. We can view responses as the means to modify geometry in the vicinity of the system. What logically follows is that object and scene geometry is in fact represented as invariant combinations of objects and responses.

Relative position is a modular, scaled property, which is probably uni-directional. In addition, it is directly related to a particular movement. Relative position is directly related to a particular displacement, in the sequential representation of things. There is also a simultaneous parallel representation of things. In our model terms, it implies the shunting linkage between the two nodes, without an external action.

Geometry, position and response are all relative properties, which are defined by a description which is valid within a window of a certain size. This window corresponds to the band pass representation of some property filter.

4 Representation in Linked Structures

The conventional way to represent association in neural network methods is to use a covariance matrix. There are however some disadvantages with such a matrix structure for the representation:

- The matrix structure and size has to be determined before the learning process starts
- It is a centralized representation, which in turn assumes a centralized computational unit
- To sufficiently well define the matrix and its components, generally requires a large number of training samples
- A conventional covariance matrix does track the relation between points mapped, but it does not track typical dynamical sequences
- There will generally be a large number of zeroes for undefined relations

As a consequence, a closed form matrix organization is not feasible for self-organizing, extendable representation structures.

Rather, an attractive representation should be oriented towards *sparse representation*, and not be organized in terms of spatial coordinates, nor in terms of feature coordinates. It is also postulated that a fundamental property of effective representation structures is the ability of *effective representation of instances*. Many procedures in neural network methodology require thousands of training runs for very simple problems. Often, there is no apparent reason for this slow learning, except that the organisation of the learning does not take into account the dynamics of the process, and considers every data point as an isolated event. We know that biological systems often require only one or two examples for learning per item. The reason is that the succession of states is a very important restrictive mechanism for compact representation as well as fast learning. The system must consequently be able to learn from single or few instances as a base for the knowledge acquisition.

As a consequence of the preceding, it is postulated that it is more important to keep track of transitions between association states, than the actual association states themselves as static points.

For that reason it is postulated that the basis of the representation is one-dimensional trajectories linking these associated states, which have earlier been referred to as canonical points. The essential properties of *similarity* or *distance* can be represented by linkages implemented by operators stating this distance in a general form. The use of discrete points is a way to resolve the problem of scaling, in that it allows the linkage of items, regardless of distance between the objects in the external physical space. The canonical points are linked by a trajectory, which is basically one-dimensional, but may fork into alternative trajectories. The linkage can be given an intuitive representation according to Figure 9.

Two or more canonical points linked together can itself be represented by a canonical point in the actual resolution and the actual set of features. It can be

Fig. 9. Intuitive illustration of linkage structure

viewed as an interval over which only one thing is happening at the level under consideration. It can, however, also be an aggregate of, or a sequence of canonical points at some level. We can view such a path as a single canonical point at a low resolution. As an example, we can take the walking along a corridor between two major landmarks. Between these two major landmarks there may well be other minor landmarks, but at a different level of resolution.

Another good argument for the representation of structures as locally one-dimensional trajectories is the *The Curse of Multi-dimensionality* [6]. This states that objects or events can not be unambiguously ordered in a space of dimensionality two or higher. The ordering sequence in multi-dimensional spaces is consequently either dependent upon rather arbitrary conventions set up, or they can be provided by contextual restrictions inferred by the spatial arrangements of items in the multi-dimensional space. This "curse" is of fundamental importance, and it has a number of effects in related fields. It is the reason that polynomials of two or more variables cannot, in general be factored. It is the reason why stability in recursive filters is very difficult to determine for filters with a dimensionality of two or higher. It is finally the reason why grammars or rule-based structures are inherently one-dimensional, although their effects may span a higher dimensionality as a result of combination, concatenation or symmetric application. This is what relates to the original issue.

An experimental system has been built up to test response driven learning of invariant structures using channel representation, with successful results [23].

4.1 Representation of Time

A crucial issue is how time should be represented in a percept/response hierarchy. The first likely choice would be to employ delay units of different magnitudes. This is a probable mechanism for low-level processing of motion. To use such delays at higher levels, implementing long time delays, has a number of problems as will appear in the ensuing discussion.

We postulate that:

Time is represented only on the response side of the pyramid. Time is in fact represented by responses, as time and dynamics are always related to a particular action of physical motion. The linkage which is related to a particular time duration is mediated by actuator control signals expressing this duration. This allows long time intervals to be implemented as the duration of complex response actions.

This gives us a consistent representation of displacement and of time. Time must be given a representation which is not a time sequence signal, but allows us to treat time like any other linked variable in the system. This is e.g. necessary as time sequence processes are to be compared. The model obtained is completely independent of the parameter scaling which generated the model. As there is not always a correspondence in time between percepts and the responses which should result, the equivalence relation must contain time as a link, rather than to match for equivalence or coincidence between the percept and the response for every time unit.

An important property of this representation is that it allows us to generate *predictive models* which allow simulations of actions in faster than real time. It is postulated that this is implemented as a direct shunting of response control signals, replacing those normally produced at the completion of a response action. See Figure 10. It is well known that there are such on-off shunts for output response signals in the nervous system, which are activated e.g. during dreaming. It is also believed that memory sequences can be processed at a much higher speed than real time, e.g. as they are consolidated into long term memory during REM sleep.

Another benefit is that something which is learned as a time sequential process, can later be represented as a completely parallel, time-delay independent model.

It appears that such an organization procedure goes in two steps:

1. The *establishment* of a model employs knowledge about the band pass adjacency between different features to build a model having the appropriate structure.
2. The *use* of a model assumes that features input to the model will exhibit the same adjacency pattern as before, although it is not tested for.

Fig. 10. Time representation for fast predictive modeling

The fact that adjacency is not crucial in the second case implies that a time sequential activation of features, in the same way as in the learning process, is no longer necessary to activate the model. Features can in fact be applied in parallel. While responses are inherently time sequential signals, we can still represent them in a time independent form as described earlier. This implies that we can activate model sets of responses in parallel. The execution of such response model sets then has to be implemented in some separate response sequencing unit. It appears likely that the *cerebellum* may have such a function [25].

Signals representing phenomena having dynamic attributes, such as responses, can be substituted by other signals giving equivalent effects. These equivalent signals will then take on the meaning of time, although they may be generated in a totally different fashion. A particular time delay is by no means necessary in the activation of the structure. This is why the dynamics of a movement can be translated into spatial distance. This is another illustration of the advantages of this coordinate free representation.

The preceding representation of time gives us in conclusion:

1. A flexible representation of time, which is scalable to the durations required and not dependent upon fixed delay units.

2. A representation of time, which is not itself represented as a time delay, but as a linkage like all other variables in the structure.
3. A linkage which can be routed over the response system for generation of actions, or be connected directly back for fast simulation and prediction.

5 The Extended Percept-Response Pyramid

The ultimate purpose of vision, or in fact all aspects of information processing, is to produce a response, be it immediate or delayed. The delayed variety includes all aspects of knowledge acquisition. This response can be the actuation of a mechanical arm to move an object from one place to another. The system can move from one environment to another. It can be the identification of an object of interest with reference to the input image, a procedure we customarily denote classification. Another example is enhancement of an image, where the system response acts upon the input image (or a copy of it) in order to modify it according to the results of an analysis. In this case, the input image or the copy is a part of the external world with respect to the system, upon which the system can act.

A major problem in the implementation of an effective vision structure is that the channel between the analysis and the response generation parts traditionally is very narrow. This implies that the information available from the analysis stage is not sufficiently rich to allow the definition of a sufficiently complex response required for a complex situation. It has also become increasingly apparent that perception cannot be treated in isolation from the response generation, firstly because a very high degree of integration is required between different levels of percepts and corresponding response primitives. Secondly, it turns out that the response to be produced at a given instance is as much dependent upon the state of the system, as the percepts impinging upon the system. The state of the system is in consequence the combination of the responses produced and the percepts associated with these responses. Thirdly, it has emerged that many classical aspects of perception, such as geometry, probably do not belong to the percept domain of a Vision system, but to the response domain.

In view of this, we want to propose a different conceptual structure, which has the potential of producing more complex responses, due to a close integration between visual interpretation and response generation [5], as illustrated in Figure 11.

This structure is an extension of the computing structure for vision, which we have developed over the years [6]. As discussed earlier, the input information enters the system at the bottom of the processing pyramid, on the left. The interpretation of the stylized Figure 11 is that components of an input are processed through a number of levels, producing features of different levels of abstraction. These percept features of different levels, generated on the left hand side of the pyramid, are brought over onto the right hand side, where they are assembled into responses, which propagate downward, and ultimately emerge at the bottom on the right hand side. A response initiative is likely to emerge at a high level,

Fig. 11. A stylized analysis-response structure viewed as a pyramid .

from where it progresses downward, through stages of step-by-step definition. This is illustrated intuitively as percepts being processed and combined until they are “reflected” back and turned into emerging responses.

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.

A characteristic feature of this structure is that the output produced from the system leaves the pyramid at the same lowest level as the input. This arrangement has particular reasons. We believe that processing on the percept side going upward in the pyramid, usually contains differentiating operations

upon data which is a mixture between input space and property space. This means that variables in the hierarchical structure may not correspond to anything which we recognise at our own conscious level as objects or events. In the generation of responses on the right hand side, information of some such abstract form is propagated downward, usually through integrating operations. Only as the emerging responses reach the interface of the system to the external world, do they have a form which is in terms of objects as we know them. In conclusion, this is the only level at which external phenomena make sense to the system; be it input or output.

This mechanism has far-reaching consequences concerning programming versus learning for intelligent systems. Information can not be "pushed" directly into the system at a higher level, it must have the correct representation for this particular level, or it will be incomprehensible to the system. A more serious problem, is that new information will have to be related to old information, on terms set by the system and organized by the system. It will require the establishment of all attribute links and contextual links, which in fact define the meaning of the introduced item. It is apparent that information can only be input to a system through the ordinary channels at the lowest level of a feature hierarchy system. Otherwise it cannot be recognized and organized in association with responses and other contextual attributes, which makes it usable for the system.

In biological systems, there appear to be levels of abstraction in the response generation system as well, such that responses are built up in steps over a number of levels [20,26]. Arguments can be made for the advantage of fragmentation of response generation models, to allow the models to be shared between different response modes.

The function of the interior of the response part of the pyramid in Figure 11 can be viewed as a more general response action command entering from the top of the structure. This command is then modified by processed percept data input entering from lower levels, to produce a more context specific response command. This is in turn made even more specific using more local, processed lower-level input data.

A typical response situation may be to stretch out the hand towards an object to grasp it. The first part of the movement is made at high speed and low precision until the hand approaches the object. Then the system goes into a mode where it compares visually the position of the hand with that of the object, and sends out correcting muscle signals to servo in on the object. The grasping response can now start, and force is applied until the pressure sensors react. After this, the object can be moved, etc.

There is a great deal of evidence that this type of hierarchical arrangement is present also at higher levels of the cortex, where a response command is modified and made specific to the contextual situation present. The processing of the cerebellar structure performs some such coordinating, context sensitive response modification [25,3]. The structure discussed seems capable of building up sufficiently complex and data-driven responses.

So far the discussion may have implied that we would have a sharp division between a percept side and a response side in the pyramid. This is certainly not the case. There will be a continuous mixture of percept and response components to various degrees in the pyramid. 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.

Acknowledgements

The author wants to acknowledge the financial support of WITAS: The Wallenberg Laboratory for Information Technology and Autonomous Systems, as well as the Swedish National Board of Technical Development. These organisations have supported a great deal of the local research and documentation work mentioned in this overview. Considerable credit should be given to the staff of the Computer Vision Laboratory of Linköping University, for discussion of the contents as well as for text and figure contributions to different parts of the manuscript.

References

1. D. Beymer and T. Poggio. Image Representations for Visual Learning. *Science*, 272:1905–1909, June 1996.
2. D. Gabor. Theory of communication. *J. Inst. Elec. Eng.*, 93(26):429–457, 1946.
3. J.-H. Gao, L. M. Parsons, J. M. Bower, J. Xiong, J. Li, and P. T. Fox. Cerebellum Implicated in Sensory Acquisition and Discrimination Rather Than Motor Control. *Science*, 272:545–547, April 1996.
4. G. H. Granlund. In search of a general picture processing operator. *Computer Graphics and Image Processing*, 8(2):155–178, 1978.
5. G. H. Granlund. Integrated analysis-response structures for robotics systems. Report LiTH-ISY-I-0932, Computer Vision Laboratory, Linköping University, Sweden, 1988.
6. G. H. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer Academic Publishers, 1995. ISBN 0-7923-9530-1.

7. W. E. L. Grimson. *Object Recognition by Computer: The Role of Geometric Constraints*. MIT Press, Cambridge, MA. USA, 1990.
8. L. Haglund, H. Knutsson, and G. H. Granlund. Scale and Orientation Adaptive Filtering. In *Proceedings of the 8th Scandinavian Conference on Image Analysis*, Tromsø, Norway, May 1993. NOBIM. Report LiTH-ISY-I-1527, Linköping University.
9. R. Held and A. Hein. Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, 56(5):872–876, October 1963.
10. R. I. G. Hughes. *The structure and interpretation of quantum mechanics*. Harvard University Press, 1989. ISBN: 0-674-84391-6.
11. L. Jacobsson and H. Wechsler. A paradigm for invariant object recognition of brightness, optical flow and binocular disparity images. *Pattern Recognition Letters*, 1:61–68, October 1982.
12. K. Kanatani. Camera rotation invariance of image characteristics. *Computer Vision, Graphics and Image Processing*, 39(3):328–354, Sept. 1987.
13. L. C. Katz and C. J. Shatz. Synaptic activity and the construction of cortical circuits. *Science*, 274:1133–1138, November 15 1996.
14. J. J. Koenderink and A. J. van Doorn. Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer. *Opt. Acta* 22, pages 773–791, 1975.
15. J. J. Koenderink and A. J. van Doorn. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.
16. T. Landelius. Behavior Representation by Growing a Learning Tree, September 1993. Thesis No. 397, ISBN 91-7871-166-5.
17. T. Landelius and H. Knutsson. A Dynamic Tree Structure for Incremental Reinforcement Learning of Good Behavior. Report LiTH-ISY-R-1628, Computer Vision Laboratory, S-581 83 Linköping, Sweden, 1994.
18. T. Landelius and H. Knutsson. Behaviorism and Reinforcement Learning. In *Proceedings, 2nd Swedish Conference on Connectionism*, pages 259–270, Skövde, March 1995.
19. T. Landelius and H. Knutsson. Reinforcement Learning Adaptive Control and Explicit Criterion Maximization. Report LiTH-ISY-R-1829, Computer Vision Laboratory, S-581 83 Linköping, Sweden, April 1996.
20. R. A. Lewitt. *Physiological Psychology*. Holt, Rinehart and Winston, 1981.
21. L. M. Lifshitz. Image segmentation via multiresolution extrema following. Tech. Report 87-012, University of North Carolina, 1987.
22. J. L. Mundy and A. Zisserman, editors. *Geometric Invariance in Computer Vision*. The MIT Press, Cambridge, MA. USA, 1992. ISBN 0-262-13285-0.
23. K. Nordberg, G. Granlund, and H. Knutsson. Representation and Learning of Invariance. In *Proceedings of IEEE International Conference on Image Processing*, Austin, Texas, November 1994. IEEE.
24. T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.
25. J. L. Raymond, S. G. Lisberger, and M. D. Mauk. The Cerebellum: A Neuronal Learning Machine? *Science*, 272:1126–1131, May 1996.
26. G. M. Shepherd. *The Synaptic Organization of the Brain*. Oxford University Press, 2nd edition, 1979.
27. S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006, 1991.