ELSEVIER

Signal Processing 74 (1999) 101–126

SIGNAL PROCESSING

The complexity of vision

Gösta H. Granlund*

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

Received 22 June 1997; received in revised form 10 May 1998

Abstract

There is no indication that it will ever be possible to find some simple trick that miraculously solves most problems in vision. It turns out that the processing system must be able to implement a model structure, the complexity of which is directly related to the structural complexity of the problem under consideration in the external world. It has become increasingly apparent that Vision cannot be treated in isolation from the response generation, because a very high degree of integration is required between different levels of percepts and corresponding response primitives. 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. In addition, 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. This article will focus on what are considered crucial problems in Vision for robotics for the future, rather than on the classical solutions today. It will discuss hierarchical architectures for combination of percept and response primitives. It will discuss 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 abstractly 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. The question of object representation is central to the paper. Equivalence is established between the representations of response, geometry and time. Finally an integrated percept–response structure is proposed for flexible response control.

Zusammenfassung

Es gibt keine Anzeichen, dass es jemals möglich sein wird, einen einfachen Trick zu finden, der wie durch ein Wunder die meisten Probleme des Sehvermögens löst. Es hat sich herausgestellt, dass das Verarbeitungssystem in der Lage sein muß, eine Modellstruktur zu implementieren, dessen Komplexität direkt von der strukturellen Komplexität des in der äußeren Welt zugrunde liegenden Problems abhängt. Es ist immer klarer geworden, dass das Sehvermögen nicht isoliert von der Antwortgenerierung betrachtet werden kann, da ein hoher Integrationsgrad zwischen verschiedenen Stufen von Empfindungen und der korrespondierenden Basisantworten benötigt wird. Wenn Empfindungen auf das System treffen, hängt die zu generierende Antwort auf die Instanz stark vom Status des Systems ab. Zusätzlich wurde klar, dass viele klassische Aspekte der Wahrnehmung, wie z.B. eine Geometrie wahrnehmen, vermutlich nicht zur Empfindungsebene eines Sehsystems sondern zur Antwortebene gehören. Dieser Artikel konzentriert sich auf diejenigen

* Corresponding author. Tel.: + 46 13 281303; fax: + 46 13 138526; e-mail: gegran@isy.liu.se

0165-1684/99/$ – see front matter
PII: S0165-1684(98)00204-7

entscheidenden Probleme, die in Zukunft für das Sehvermögen von Robotern von Bedeutung sind und nicht auf die klassischen Probleme der Gegenwart. Er diskutiert hierachische Architekturen zur Kombination von Empfindungen und Basisantworten. Er diskutiert das Konzept kombinierter Invarianzen zwischen Empfindung und Antwort als wichtiges strukturelles Element des Sehvermögens. Beibehalten wird, dass Lernen essentiell ist, um die notwendige Flexibilität und Anpassungsfähigkeit zu erhalten. Als Konsequenz wird argumentiert, dass Invarianzen für den Zweck des Sehvermögens nicht abstrakt geometrisch sind, sondern aus der Interaktion zwischen Empfindung/Antwort mit der Umgebung folgen. Die Frage der Informationsdarstellung wird extrem wichtig in verteilten Strukturen der vorgesehenen Typen, in denen unsichere Information zur Aktualisierung von Modellen und assoziierten Daten verwendet wird. Die Frage der Objektdarstellung ist ein zentrales Thema des Artikels, in dem auch Gleichwertigkeit zwischen der Darstellung von Antwort, Geometrie und Zeit hergestellt wird. Schließlich wird eine integrierte Struktur zwischen Empfindung und Antwort für eine flexible Antwortkontrolle vorgeschlagen.

Résumé

Il n'y a pas d'indication qu'il sera un jour possible de trouver un truc simple qui résoudra miraculeusement la plupart des problèmes de vision. Il apparaît que le système de traitement doit être capable d'implémenter une structure modèle, dont la complexité est directement liée à la complexité structurelle du problème considéré dans le monde extérieur. Il est devenu de plus en plus apparent que la vision ne peut être traitée de façon isolée de la génération de réponse, parce qu'un très haut degré d'intégration est nécessaire entre les différents niveaux de perception et les primitives de réponses correspondantes. La réponse à produire à une instance dépend autant de l'état du système que des perceptions affectant sur le système. De plus, il est devenu clair que de nombreux aspects classiques de la perception, comme la géométrie, n'appartiennent probablement pas au domaine de perception de la vision, mais à celui de la réponse. Cet article se concentrera sur ce que nous considérons comme des problèmes cruciaux en vision pour la robotique dans le futur, plutôt que sur les solutions classiques d'aujourdhui. On y discutera des architectures hiérarchiques pour la combinaison des primitives de perception et de réponse. On y discutera le concept des invariants de perceptions–réponses combinés comme étant des éléments structurels importants de la vision. Nous y soutiendrons que l'apprentissage est essentiel pour obtenir la flexibilité et l'adaptabilité nécessaires. Par conséquent, nous argumenterons que les invariants pour le vision ne sont pas géométriques de façon abstraite mais définis à partir de l'interaction perception–réponse avec l'environnement. Le problème de la représentation de l'information devient extrêmement important dans des structures distribuées des types prévus, où l'incertitude de l'information doit être établie pour améliorer les modèles et les données associées. La question de la représentation des objets est centrale dans cet article. Une équivalence est établie entre les représentations de réponses, la géométrie et le temps. Finalement, une structure intégrée perception–réponse est proposée pour un contrôle de la réponse flexible.

Keywords: Vision; Robotics; Information representation; Hierarchies; Learning; Linkage structures; Semantic networks; Response generation

1. Introduction

There is no indication that it will ever be possible to find some 'simple trick' that miraculously solves most problems in vision. It turns out that the processing system must be able to implement a model structure, the complexity of which is directly related to the structural complexity of the problem under consideration in the external world.

The traditional methodology for vision contains many procedures to perform various tasks [2,4,14]. A common problem is that these procedures are rarely suitable as components of a larger system. The reason is that information is represented in different ways for different types of features. It is difficult to have such descriptors combine their statements in a graceful way, and to have them control operations in a parametric way.

A particular object may appear in many different orientations, sizes, projections, etc. It is necessary to deal with this variability in a more efficient manner than to have one model or template for every possible orientation of an object. It is necessary to find model representations that exhibit *invariance*, i.e. in

which the descriptive statements do not depend upon what are considered irrelevant variations such as the orientation of an object.

What is irrelevant, or an *invariant mode*, will differ from one situation to another. In one case we want to recognize an object regardless of its orientation. In another case the orientation of the object is of major concern. In a systematic approach, we will try to design models to produce different modes separately, such that we can use the information if desired.

It turns out to be necessary to use what we may call different sub-algorithms or sub-models on different parts of an image. The selection of a particular sub-algorithm is usually based upon a tentative analysis of the image content. The techniques for such preliminary or *preattentive* vision have been extensively developed [1,30,40], although they are outside the scope of this paper.

The reason for using different sub-algorithms is the simple fact that not all possible events can be expected in a particular context. As indicated earlier, this handling of sub-algorithms has to be implemented as a modification or a parameterization of more general algorithms. Without such a system of sub-algorithms, we would obtain a model computing structure which was totally unmanageable, which would exhibit a combinational explosion.

In order to allow the repeated use of existing models or model parts, we will try to design models that can be used for several purposes. These partial models will have some general properties of invariance generation and generality in representation.

Our own work, as well as some of the work cited, has received a great deal of inspiration from what is known about biological visual systems [15,16,29]. This is not to say that the mechanisms presented in the sections to follow are necessarily models of phenomena in biological visual systems. Too little is so far known to draw any firm parallels. The ultimate criterion for our interest is performance from a technical point of view.

This article will focus on what are considered crucial problems in Vision for robotics for the future, rather than on the classical solutions today. It will discuss hierarchical architectures for combination of percept and response primitives. It will discuss 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 abstractly 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. The question of object representation is central to the paper. Equivalence is established between the representations of response, geometry and time. Finally, an integrated percept–response structure is proposed for flexible response control.

2. Structured representation of feature information

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. Without such standardized information representations, the complexity will be overwhelming and the functional mechanisms completely obscure. One way to satisfy these requirements is to implement the model structure in a regular fashion. It is often useful to view this regular arrangement as a *hierarchy*, although we should bear in mind that the communication need not be restricted to adjacent layers of such a hierarchy. In principle, hierarchical structures are nothing new in information processing in general, or in computer vision in particular. A regular organization of algorithms has always been a desired goal for computer scientists. However, in order for such a structure to work effectively on spatial data, certain crucial requirements have to be fulfilled for information representation and for the structures of operations.

2.1. The feature abstraction pyramid

We can distinguish between two different types of hierarchies:

- Scale hierarchies;
- Abstraction hierarchies.

Most of the work on hierarchies 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 [9,22,28,42].

Granlund introduced an explicit abstraction hierarchy [7], employing symmetry properties implemented by Gaussian envelope functions, in what today is commonly referred to as Gabor functions or wavelets [5]. 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 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

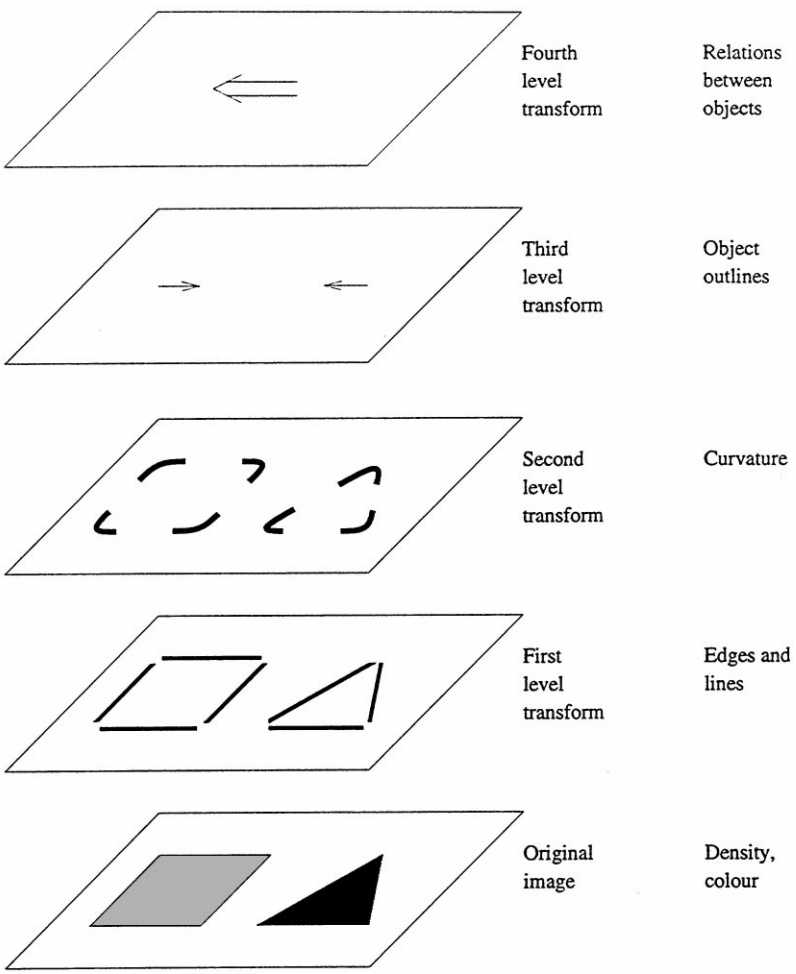

Fig. 1. Conceptual representation of an image as an abstraction hierarchy.

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.

The example in Fig. 1 gives an intuitive idea of the notion of abstraction levels. It should not be taken as the only possible arrangement of an abstraction hierarchy.

We furthermore assume a linkage between elements at different levels of abstraction. It is easy to accept the notion that line elements at an angle to each other form a convex structure element, representing curvature. Similarly, a number of convex structures will in combination infer an outline, which is a characteristic of an object.

Although the issues of size and scale, and of level of abstraction are conceptually different, they become related in the implementation. With increased level of abstraction, generally follows an increase of the scale over which we relate phenomena. In a region there may be few objects, but they may comprise a large number of lines and edges. The output of a transformation can generally be described using fewer samples than the input without any loss of information. This compression is called *sub-sampling*. A 256×256 transform image at one level is, after an operation, represented by 128×128 elements at the next level. This level-by-level decrease in the amount of data, by a factor of two in each dimension, produces a pyramid representation. For that reason we often refer to the structure as an *abstraction pyramid*. We will deal with such a combination of abstraction levels and variation of scale in the following section.

2.2. Information representation in a hierarchy

It is necessary to omit the important discussion of mathematical representations for information in this document. For further details, reference has to be made to [9]. The generation of descriptors can generally be viewed as a two stage process. See Fig. 2.

In the first stage, a number of filters are convolved with the image content within a window to produce a parameter vector. These filters can be chosen in a number of ways such as Gaussian, quadrature or any of the families of wavelets. The parameters in the vector are generally correlated to form a subspace of the original vector space. The second stage implies the forming of more compact and invariant representations. This has the purpose to reduce the amount of redundant information, but also produce descriptors which appear to represent important conceptual properties like orientation, size, etc.

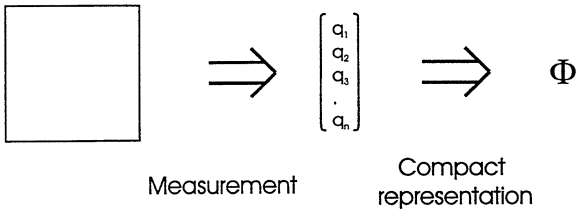

Fig. 2. Generation of a compact information representation viewed as a two-stage process.

It turns out that for 2-D information a *vector* representation is advantageous, while for three 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. They also give a description of outcomes in relation to a continuous *metric*. The effects of this will be dealt with in a later section. For further details, reference has to be made to [9], where in addition examples of processing of images are given.

2.3. Model-based, top–down processing

What we have discussed so far are the properties of processing as we go upward in the processing pyramid. This is often referred to as *bottom–up* processing. In the same way that context information affects our interpretation of a more local event, context information determines the possible alternatives of events within a local region, and consequently the operations we want to perform. The question is how 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 Fig. 3.

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 filtering in multiple dimensions [11]. This is a very important issue which, however, goes beyond the objectives of this document, and reference has to be made to [9], where in addition examples of processing of images are given.

3. Representation as linked structures versus arrays

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 attractive.

The increasing use of actively controlled and multiple sensors requires a more flexible processing and representation structure. The data which arrives from the sensor(s) can be viewed as image patches of different sizes, rather than frame data in a regular stream. 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 of 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 in a data structure.

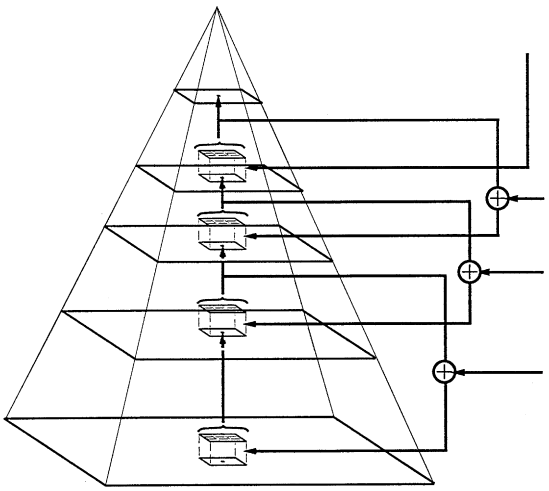

Fig. 3. A hierarchical structure with bottom–up and top–down flow of information.

The conventional iconic array form of image information is impractical as it has to be searched and processed every time some action is to be performed. It is desirable to have the information in some partly interpreted form to fulfill its purpose to rapidly evoke actions. Information in interpreted form, implies that it should be represented in terms of content or *semantic* information, rather than in terms of array values. Content and semantics implies *relations* between units of information or symbols. For that reason it is useful to represent the information as relations between objects or 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:

- An array implies a given size frame, which cannot easily be extended to incorporate a partially overlapping frame;
- Features of interest may be very sparse over parts of an array, leaving a large number of unused positions in the array;
- A description of additional detail cannot easily be added to a particular part of an array.

3.1. The feature array as a linked structure

How do we go from the hierarchical analysis and description of a patch in the form of an array, to the representation in a linked structure? Let us assume that we have generated arrays containing the original image, the orientation description and the curvature description, according to the lowest abstraction levels indicated in Fig. 1 [9].

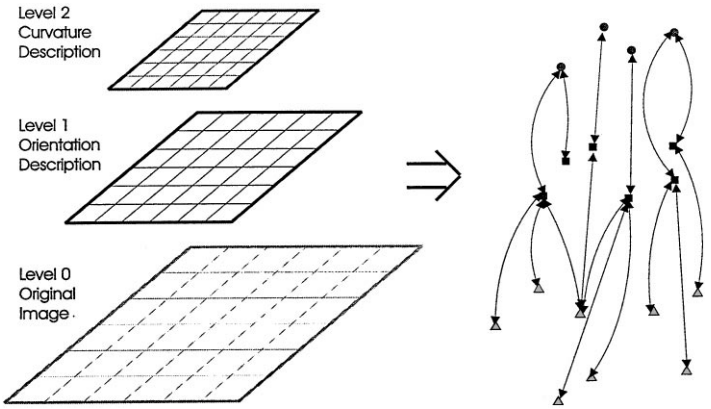

Fig. 4. Transition between hierarchical array structure and linkage structure.

These descriptions reappear in the array pyramid on the left-hand side of Fig. 4. It is possible to build up a set of maps or lists, where each set of maps contains elements of a particular abstraction level, and where each element contains pointers to elements at a lower level. See the right-hand side of Fig. 4, which is an intuitive illustration of the fact that we no longer have a regular array structure which relates elements. In this illustration, no particular attempt has been made to distinguish between feature links and relative position links. That distinction will be dealt with later.

In this way a linked structure is built up, which represents the object in question. What is particular for this linked structure is that the maps or lists arrange the objects in terms of reference, or in semantic terms. In this structure, primitives are tied together at higher level primitives. The object is represented as a free-floating structure of linked primitives, or a *graph* [12]. This structure or graph can now be matched to other graphs for recognition, or related to a larger structure, of which it may be a part. For the case of a general graph, we know that this is not a trivial task.

3.2. The classical problem of graph interpretation

Hierarchical features derived from an abstraction pyramid can be used for classification of objects and scenes in the classical way by putting labels on them. For more complex structures, an extension of the labeling procedure is to express how the differently labeled objects are related in the form of a *labeled graph*. This graph is then subjected to an interpretation.

A great deal of computational power is required for a procedure where the graph is generated first, followed by an interpretation of this graph. The reason is that it is more difficult to analyze a given graph, than to deal with the original problem structure in question, with its attributes available. It turns out that the structure graph becomes less specific, as it is made more abstract and not related to particular inputs.

In addition, the label interpretation substructure requires a centralized decision structure, which requires an overview not only of this particular graph structure, but other related graph structures which may emerge as well. This would seem to preclude the use of decentralized functions in a system, an organization which is deemed necessary for a system which can self-organize its information storage and computations, and learn. We will in the subsequent sections see how we can make the graph provide its own outputs.

3.3. A linked feature hierarchy

How can a better structure be devised, which does not require an external system for interpretation? We assume that we are looking at a single feature level, say orientation, in the present

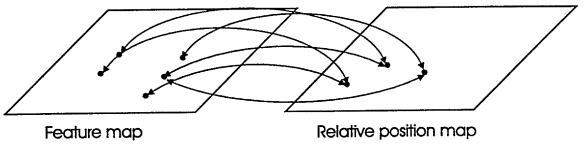

Fig. 5. Generation of continuous maps of features and of relative position, and the establishment of links between these to describe objects.

discussion. We can view Fig. 5 as an attempt towards a linkage structure for object representation at a particular level.

The transform level of Fig. 1 under consideration, is scanned for features above a certain threshold. Such instances are mapped into a list, or rather a *feature map*, which now represents the features ordered continuously according to *similarity*. See Fig. 5, left side. The issue of similarity or adjacency is crucial, and will be discussed extensively in the following sections. A certain map has a particular locality scale, within which features are mapped without regard to other properties such as position. As an example in the cortex, orientation is mapped continuously into regions or "blobs" of a certain size.

Contributions within a particular level are located, and are assigned to an element in the map, depending upon the class membership of the feature. The system does as well keep track of their position. For conventional array descriptions we use geometric Euclidean distance to describe the distance between two objects as a numerical parameter. For reasons which will become apparent, we want to avoid free floating labels of properties, be they local features such as color, orientation, etc. or geometric properties such as distance, size, etc. We want to refer all such properties to a common reference or *map*. The relative position between features in the feature map is represented by links to the relative position map, as indicated in Fig. 5, right side.

In such a way, the feature array is broken up into two different parts: a feature map and a relative position map, both of which are arranged with respect to similarity or property distance. The directly geometrical distance relations between items in an array are replaced by references to a relative position map, describing this geometry.

What is the advantage of breaking up the seemingly well-behaved geometrical object description into bits and pieces of different types? The most important reason is that it provides the basis of *invariance* mechanisms of different types.

A particular set of features can be linked by several different relative position references, thereby ensuring e.g. a scale invariance of the object description. This can be viewed as the most primitive version of a WHAT and WHERE system division, something which is known to take place in the primate visual system [31]. These mechanisms will be further developed and discussed in the ensuing sections. The second, very important reason is that we will see how we can substitute geometry with response links, which miraculously gives us the required links from the network out into the external world.

In addition to the lateral linkage, there is a vertical linkage between levels. We saw in the discussion around Fig. 4, that features at a lower level combine into more complex, higher level features. Two line elements at an angle will comprise a corner, etc. Higher level features are assigned to a different list or map, which contains pointers to the lower level feature elements and relative positions building up the element in question. There are as well links over the relative position or displacement map, valid for objects at this level. See Fig. 6.

We will according to Fig. 6, have different object level sets of maps; each level containing two complementary maps:

- A feature map;
- A relative position (response) map.

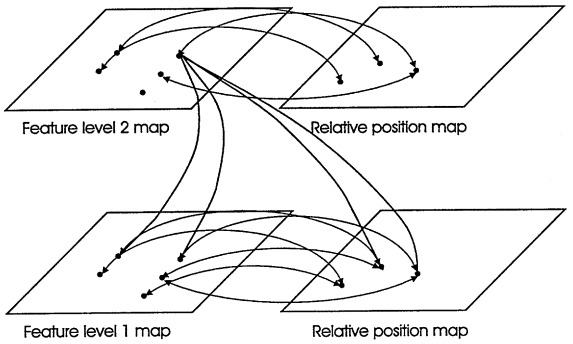

Fig. 6. Generation of maps of features and of relative position at different levels, and the establishment of links between these to describe objects.

There will be several different families of maps, which will have different purposes in producing a complete and convergent representation. Localized, relative position references deal with the invariant representation of objects, or the WHAT system. Absolute position references are necessary for reaching and for the manipulation of objects, which gives the WHERE system [31].

3.4. Output from linked networks

From the preceding discussion, it is apparent that it is necessary to build up a network structure, which includes outputs as well as inputs for the information. This implies that a response linkage structure has to be built up simultaneously with a percept interpretation and linkage structure. We will see in the next section how a network organization driven by responses can solve these problems for us.

4. 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 *mode association*; to ascertain that there are responses which are associated with a particular percept or percept transformation.
   - The second reason is one of simplification, 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 from the percept structure, 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 [13]. 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.

Driving a learning system using response signals for organization, is a well-known phenomenon 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 [20].

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 repertoir, 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.

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 [23–26].

5. Object representation using percept–response invariants

Over the years there has been an increasing interest in research on invariants [18,19,21,32]. 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 undoubtedly 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.

An important application of invariant representation is for object description. To position ourselves for a thorough analysis, we will look at two traditional major lines of approach which have been used for object description: *object-centered* and *view-centered* representation. See Fig. 7.

From the real object, a number of measurements or projections are produced. See Fig. 7(a). 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 [10]. See Fig. 7(b). The image appearance of an instance of a particular orientation of the object is then obtained using separate projection mappings.

A view-centered representation, on the other hand, combines a set of appearances of an object, without trying to make any closed-form representation [3,34,39]. See Fig. 7(c).

5.1. Object-centered representation

The basic idea of the object-centered representation is to produce a representation which is as

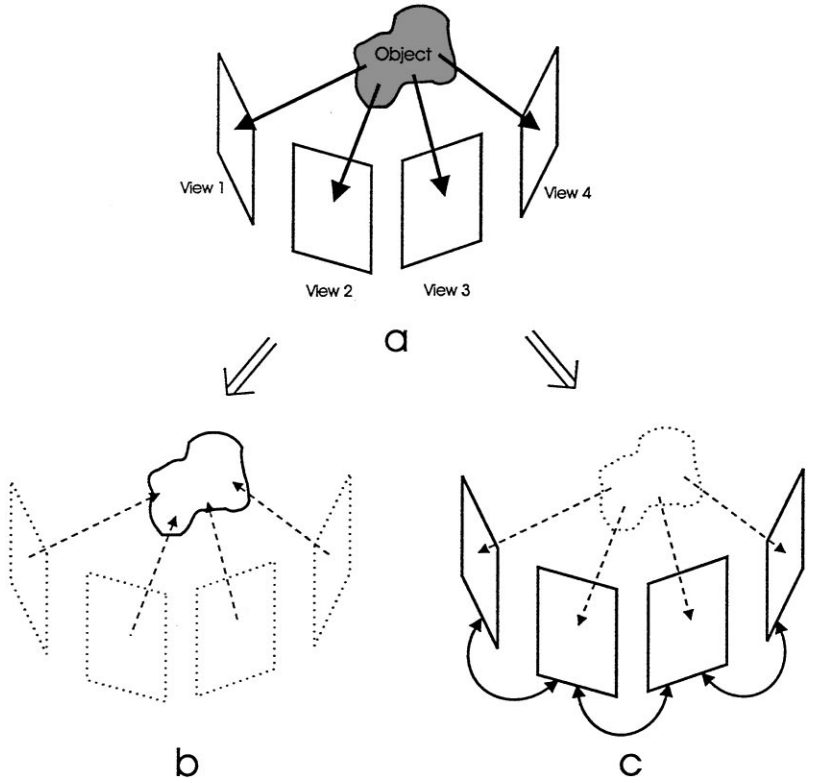

Fig. 7. Object-centered and view-centered representation of an object. (a) Measurements produce information about different views or aspects of an object. (b) Object-centered representation: The views are used to reconstruct a closed form object representation. (c) View-centered representation: The views are retained as entities which linked together form a representation of the object.

compact and as invariant as possible. It generally produces a closed-form representation, which can 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 object should be done more easily as the object description is independent of any 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 [9], as a simple form of object-centered representations. 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 subspace. 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 [9].

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 response domain procedural representation.

5.2. View-centered representation

In a view-linked 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 responses which caused the particular view. 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. This property will be shown to be crucial for the development of a learning percept–response structure. A view-linked representation in addition, has the advantage of being potentially self-organizing. There are also indications from perceptual experiments, which support the view-centered representation.

An important reason for the view representation is that we want an interpretation, rather than a copy of an object that we want to deal with. It provides us with links to the response side of the system as we require.

An object-centered representation is by definition normalized 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 we have with specific contexts or response procedures which are related to that context or view.

A normalization, or the generation of an invariant representation, implies discarding information which may be essential for the system to act using the information. It is important to remember that we are not interested in objects, but in situations and generation of responses.

5.3. Combination of representation properties

It is postulated that we can represent objects as invariant combinations of percepts and responses. This suggests that we shall start out from the view-centered representation of objects, and interpret this in the light of invariant combinations of percepts and responses. As we will see, certain aspects of generalization are necessary, which will bring us to something which shares important properties of the two extreme varieties. As an example, an efficient representation of instances is necessary for a representation structure.

The structure which results from the preceding model will be of type frames-within-frames, where individual transformations of separate objects are possible within a larger scene frame. See Fig. 6. This seems 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 discussed elsewhere. In this way, the response map 'reaches' into the frame in question, to implement the percept–response invariance of a particular object aspect.

5.4. Object properties being part percept and part response

Vision has traditionally 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 [35].

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.

A plane is not a distinguishable entity in the percept domain. 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, which is not located in the percept domain but in the response domain.

Similarly, it is believed that projective transformations are to a large extent response domain features. They describe how the external world changes its appearance to the system as a function of our movements in it. The primary step in that modeling is to relate the transformations to egomotion. The secondary step is for the system to generalize and relate the transformation 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 them to any situation of relative motion of the object in question.

In the same way, the representation of *time* is postulated to be residing on the response side of the structure. A further discussion of this is given in Section 8.1.

5.5. Extension of view linkage

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.

We postulate that the invariant property is the combination of percept structure and response structure. There are different ways to interpret this combination of views and response states to form an object:

View + Change in position = Invariant;

View + View Linkage = Invariant;

View + View Linkage = Object.

The combination of views with information concerning the position of these views, which is equivalent to the combination of percepts and responses, will in return allow an interpretation for any required angle 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. This model goes well with the notion that we may not necessarily know how an object appears from all views or sides. We can never expect completeness in this respect.

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 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 useful 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. This also implies that the observation of a particular set of percepts matches or infers a particular state of the system in terms of responses, or what we commonly denote context. To emphasize, they are equivalent precisely because the combination of them forms an invariant; an entity whose variation is not perceivable in some 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.

5.6. 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 features and responses.

Relative position is a modular, scaled property, which is uni-modal, and directly related to a particular displacement, in the sequential representation of responses. There is also a simultaneous parallel representation of response effects or geometry. In our model terms, this implies the shunting linkage between two nodes, without an action into the external world.

Geometry, position and response are all relative and local properties, which are defined within a window of a certain size. This window corresponds to the band pass representation of some property filter, as will be described in a subsequent section.

5.7. Object versus interpretation

In the process to generate appropriate responses to scenes, objects, etc., we require information describing the situation to the system in an appropriate way and with an appropriate accuracy. An interpretation of an object is a limited representation of associated states and responses, which is related to a particular contextual or observation state. As such, an interpretation or a description is something *entirely different* from the object it relates to, not just some incomplete version of the object. In literature there is sometimes a misunderstanding expressed, in that we want something which is as exact a copy of an object as possible. This is not true because a copy is an uninterpreted version, which does not help us. The ultimately absurd consequence of this is, the sometime stated view, that the best representation of an object is the object itself.

We can get some insight into the issues, borrowing the view in Quantum Mechanics, where an objective, unrestricted, uninterpreted field world is subjected to interpretation or observation through the projection on a Hermitian operator, which provides a local, limited description [17]. See Fig. 8. It is important to understand that the object and the description domains are entirely different things, not just a matter of fidelity.

It appears that what we require in the spatial-cognitive part of the system is a representation of what we may call *situations* rather than objects. If we find a telephone up-side down on our desk, it is a different situation than if we find it in a normal position, as it requires a different set of responses to get it into operation and dial a number. This is the type of situation that the system has to be able to deal with. The fact that we are dealing with the same object, something called a telephone, does not help particularly. We have earlier challenged the view that the best representation of an object is the object itself. To carry around the object itself, or a copy of it, does not help the system to do anything as it is not in a form suitable for action.

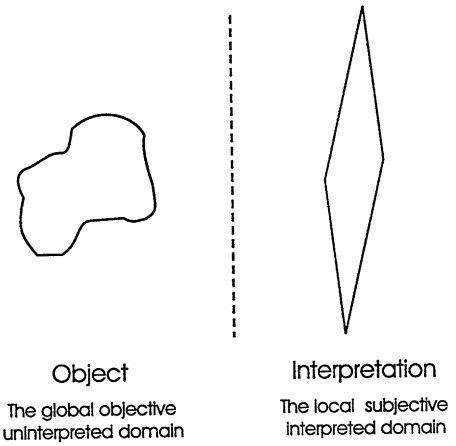

Fig. 8. Illustration of the two distinctly different domains of object versus description or interpretation.

An object appearing under a particular view is consequently something different from its appearance under another view, as they are related to different response situations. This means that there is probably no such thing as a standard view of an object, although there may be a most commonly occurring view.

5.8. What do we know about vision, or The peril of inference from conscious experience

As in all aspects of technology, we use our minds to devise methods for vision and intelligent systems. This is the only mode of operation available, but in the case of spatial cognitive information processing, we have to be particularly cautious.

A traditional goal in all aspects of information processing is to establish a class membership identity of a signal; to denote a particular object. The class memberships that we use for such a labeling are derived from our conscious experience of interaction with objects. Similarly, we tend to break up these objects into parts, which are themselves consciously recognizable objects or entities.

A potential danger inherent in this process, is 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. This is in psychology and in Artificial Intelligence referred to as two distinctly different domains of processing and experience: *procedural* and *declarative*. The *procedural* domain deals with the development of motor skills and highly contextually dependent experiences, which may be difficult to express in language. A typical example is to describe how to ride a bicycle. The *declarative* domain deals with facts and events which we can handle consciously and describe verbally or communicate. 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. This has the consequence that

> *Our conscious perception of the external world is in terms of the actions we can perform upon objects around us*

rather than some window through our eyes out into the environment. There is a great deal of perception research, which supports this notion. Experiments have been performed on so called split-brain subjects, where the *corpus callosum*, which communicates information between the two hemispheres of the brain is either sectioned or non-existent from birth. Experiments can be set up where information has to be communicated indirectly and externally from the cognitive half to the motor/language/consciousness half of the brain. The conscious part can then be deceived about what has really happened, but it will still try to produce as plausible a conscious experience and rationalization as possible [6,37,38]. With reference to the pyramid structure in Fig. 11, we may with some exaggeration say that the motor/language/consciousness is just the lower right part of this pyramid.

This implies that 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 objects in conscious terms emerge, and are available for motor manipulation. Somehow, we are too late to see *how* it happened. We only notice *what* has happened.

5.8.1. 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. What we have been talking about above is 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 anyway not be in exactly the same state as the sender of the message. 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 have 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 due to the rich structure that it evokes in the receiver's cognitive system; not due to the structure of the sentence itself. Most of the information necessary for the response has to be contained in the structure of the receiver; it cannot 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.

6. 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 or filter 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.

In robotics we have traditionally assumed a structure according to Fig. 9, where a vision system is controlling an actuator system. The systems were viewed as sophisticated perception modules, to which a few actuator wires were attached, causing the requested responses.

A major problem in the implementation of such a system structure is that the channel between the analysis and the response generation parts 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 become increasingly apparent that perception cannot be treated in isolation from the response generation, firstly because a very high degree of

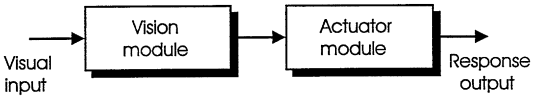

Fig. 9. Simplest structure of robotics control system.

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, do not belong to the percept domain of a Vision system, but to the response domain.

It seems that there has to be two tiers of organisation within an effective computation structure for spatial information, see Fig. 10. Within the lower tier there is an organization of data in relation to external geometry. This is also true for biological systems, where low-level orientation description is mapped upon the cortex in a accordance with position in the visual field. Similarly for motor functions and other features, which are mapped correspondingly between the body and the cortex. For technical systems, it can be assumed that computations to produce these low-level features can be made in parallel, and that influences on earlier levels of computation are at least very local.

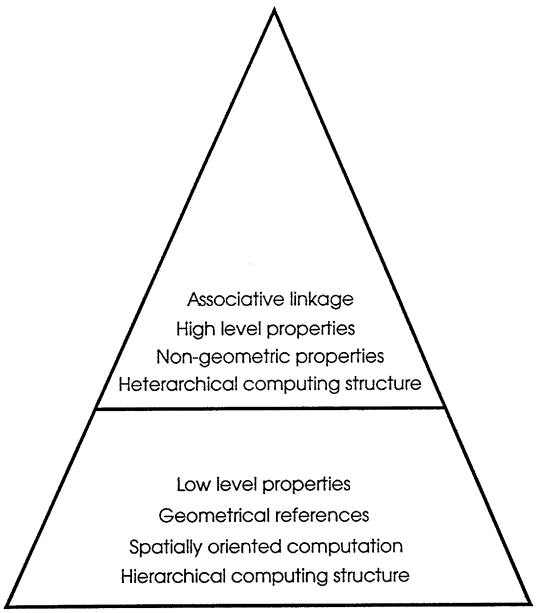

Fig. 10. The two-tier pyramid.

It is postulated that at some level of abstraction, local geometrical relations become less important, and other non-spatial and non-local relations become essential. The separation into distinctive paths for WHAT and WHERE information is one indication of this in biological systems [31]. This forms the upper tier of the computation structure for spatial information, see Fig. 10, which has to be formed through association based upon properties of the signals themselves as they are driven by stimuli, rather than by geometric adjacency. Most of the discussion to follow will deal with computation structures for the upper tier.

In contrast to the system structure in Fig. 9, we want to propose a conceptual structure, which has the potential of producing more complex responses, due to a close integration between visual interpretation and response generation [8], as illustrated in Fig. 11.

This structure is an extension of the computing structure for vision, which we have developed over the years [9]. As discussed earlier, the input information enters the system at the bottom of the processing pyramid, on the left. The interpretation of the stylized Fig. 11 is that components of an

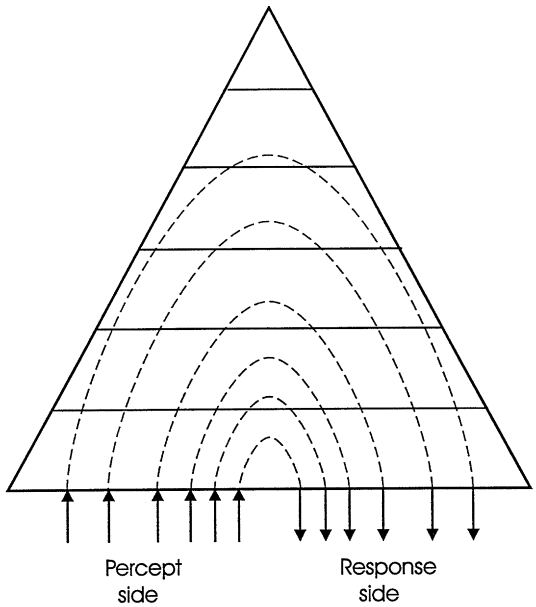

Fig. 11. A stylized analysis–response structure viewed as a pyramid.

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, 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 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. More on this in Section 7.2. This means that variables in the hierarchical structure will not correspond to anything which we recognize 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 structure illustrates far-reaching consequences concerning programming versus learning for intelligent systems. Information cannot 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, which we will deal with later, 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 [27,36]. Arguments can be made for the advantage of fragmentation of response generation models, to allow the models to be shared between different response modes.

A look into the interior of the response part of the pyramid in Fig. 11 will reveal a stylized structure for implementation of responses as can be seen in Fig. 12. We can view this as a more general response action command entering from the top of

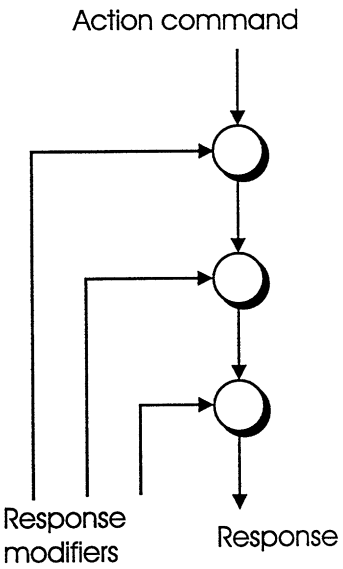

Fig. 12. Multi-level step-wise definition of a response.

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 it 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.

Interpreted in classical terms, we can state that knowledge in this pyramid is of two types:
1. Feature-related knowledge;
2. Response-related knowledge.
These two types of knowledge relate directly to the two sides of the feature–response pyramid.

It should be emphasized that there is no sharp division between a percept side and a response side in the pyramid. 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.

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.

7. Non-geometric similarity representations for linked structures

There is a great deal of literature available on the topic of object representation using classical methods [10], 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.

7.1. 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, cannot 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 the concepts of *invariance* and *equivariance*, which are discussed elsewhere in this paper.

From statistics we have derived the notion that we can only certify the *difference* between items. As a complement, we can never be certain that two items are similar. It may only be necessary to take one more measurement to be able to separate the items, or that we are able *to reject the hypothesis of similarity*. Consequently, similarity is never unequivocal. It may be rejected if we look more closely at the items, and make another measurement which happens to be distinctive. Dissimilarity on the other hand is absolute. There is no way in which we may add more measurements to make items more similar, from a statistical point of view.

In practice, statistics may not give the full picture, as structural aspects enter the setup of the

problem. Measurements which we try to relate must be commensurable, which means that we must have already identified the object and its parts under investigation. This is generally a severe catch. Measurements may be partially incommensurable, such as the two views of a stereo image, or properties at two instances of a moving object. In general, the measurement of dynamic properties leads to measurements of properties in two or more images which are not completely commensurable. This does also apply to the comparison of two different objects in general. Measurements upon two objects can never be fully commensurable as they deal with what is in effect two different things. Measurements made on one of the objects are not totally relevant with respect to the other object. This leads to a variety of the *Uncertainty Principle* in vision [41].

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 in Fig. 13.

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

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

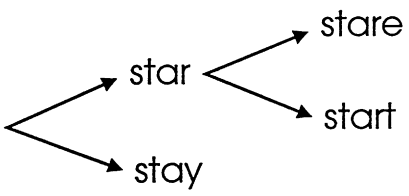

Fig. 13. Example of words having a small distance in terms of an ASCII letter metric, but large distances in content or meaning.

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.

7.2. Channel representation of information

A representation of similarity requires that we have a *metric* or 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 Fig. 14.

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. 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. If we view the channel output as derived from a band pass filter, we can establish a measure of *distance* or *similarity* in terms of the parameters of this filter. See Fig. 14. For a

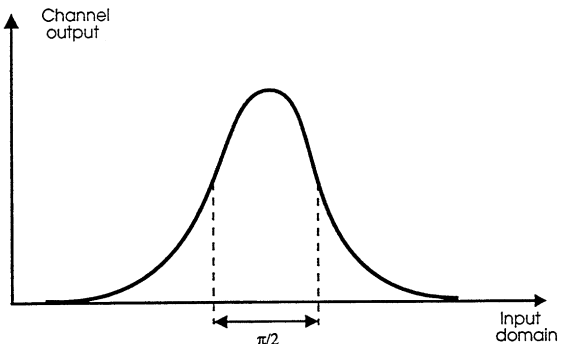

Fig. 14. Channel representation of some property as a function of match between filter and input pattern domain.

conventional, linear simple band pass filter, the phase distance between the flanks is a constant $\pi/2$. Different filters will have different bandwidths, but we can view this as a standard unit of similarity or distance, in terms of the particular channel filter. 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. The output of the channel is dependent upon the value of the input parameter relative to the center of the channel. There will be an ambiguity in terms of the position with respect to this center, which can only be resolved with the combination with other channel responses.

Our mission is to obtain a description or model of some phenomenon involving a large number of channel signal contributions. Such a description can be achieved under certain assumptions of continuity in the signal structure. A description implies that we relate a phenomenon to other phenomena, described in terms of various properties through a comparison of difference and similarity in these properties in relation to various alternatives. This procedure will, if successful, allow us to link or combine different originally independent fragments, implying that we form an abstract spatial structure for some phenomenon.

If a set of adjacent channel receptors driven by a *single and simple* stimulus display an output as indicated in Fig. 15, where the falling flank of one channel coincides with the rising flank of another channel, we know that the distance between the peaks is as well $\pi/2$. This phase distance is in terms of some average of the parameters of the channel filters involved.

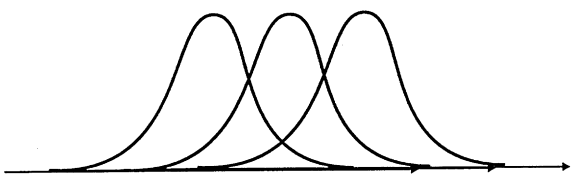

Fig. 15. Visualization of channels as partially overlapping entities in some space.

By using a sufficiently dense representation we can employ the knowledge of a particular similarity or distance between channel contributions. Viewed in the input parameter domain, we can intuitively view a sequence of activated channels as illustrated in Fig. 15. The input parameter may be time, displacement or some other variable.

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 Fig. 14, 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 is moving over the spatial region of the detector, or a line at the proper position is rotating over the detector, or in general a combination of both.

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 Fig. 16, which is intended as a two-dimensional version of Fig. 15. 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.

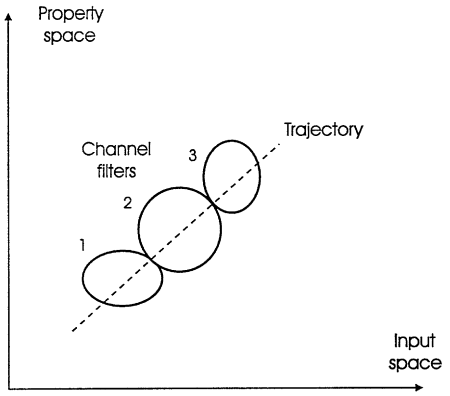

Fig. 16. Visualization of channels in input space as well as property space.

Given this fact that we are dealing with projections in different subspaces, we must concede that the illustrations in Figs. 15 and 16 are not totally correct, although hopefully helpful. 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. Each subspace can for intuitive visualization be represented as a vector, which is orthogonal to its nearest-neighbor subspaces. This is illustrated in Fig. 17. As can be seen from Fig. 17, the vector magnitudes are tapering off from the center, which is the assumed point of observation of the coordinate system. This indicates that while adjacent subspaces are orthogonal, we cannot say much about the relation to 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.

7.3. Implications of multiple measurements

From the previous section it follows that similarity is measured and valid along a single, one-

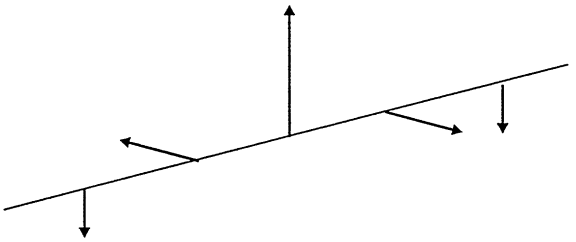

Fig. 17. Representation of channels as orthogonal subspaces.

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. Two successive links may not represent distances along the same, one-dimensional subspace, as we have no way to track what filters are involved where. There is no way to order objects unambiguously for two different reasons:

1. There is no way to order items which are defined in a multidimensional space, which is the *Curse of Multi-dimensionality* [9].
2. It is not possible to sort objects with respect to a particular property, as 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 Fig. 18.

7.4. 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 output from the filters involved. It is postulated that canonical points correspond to phases 0°, 180° and $\pm 90^\circ$ in outputs

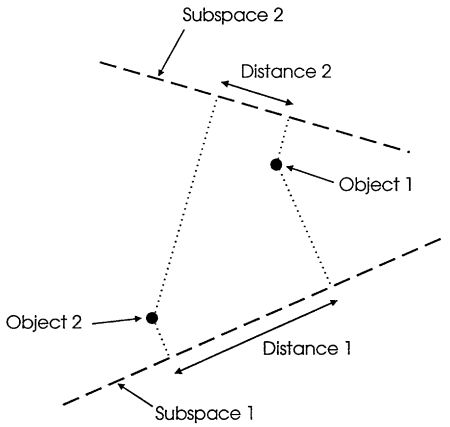

Fig. 18. Distance between two objects measured with two different probes, implying projection upon two different subspaces.

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 [17].

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 band pass function. The implementation of this is apparent for low-level features, but we can find equivalent interpretations at higher levels.

8. Associative linkage between percept and response primitives

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 assumes a centralized computational resource.
- To sufficiently well define the matrix and the 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 attractive 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 *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 organization 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 usually 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

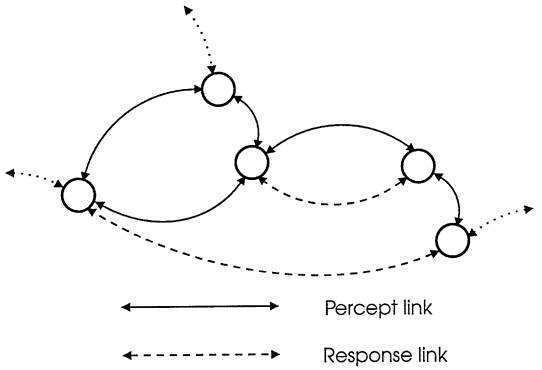

Fig. 19. Intuitive illustration of linkage structure.

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 Fig. 19.

A link between two canonical points can itself be represented by a canonical point in the actual resolution and the actual set of features. It can be 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.

An experimental system has been built up to test response driven learning of invariant structures using channel representation, with successful results [33].

8.1. Representation of time

A crucial issue is how time should be represented in a percept/response hierarchy. An apparent way is to employ delay units of different magnitudes. This is a useful 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 represented by the duration of 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 gives us a linkage structure representation of displacement and of time. Time must be given a representation which is not a time-delayed version of features, 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. In the same way that we in an earlier section postulated equivalence between response and geometry, we can now postulate that:

Response and time are equivalent

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 direct shunting of response control signals, replacing those normally produced at the completion of a response action. See Fig. 20. 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

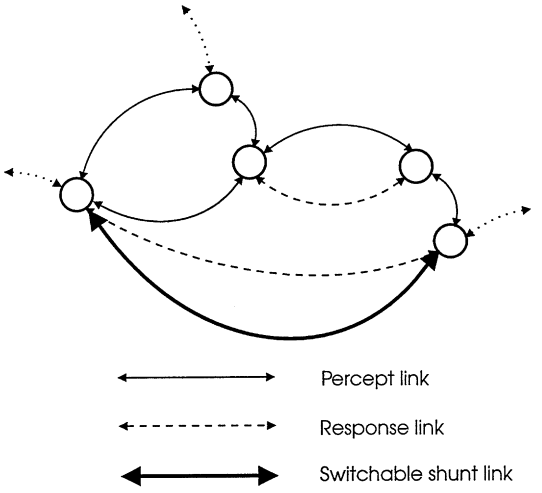

Fig. 20. Time representation for fast predictive modeling.

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 sequentially 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, which allows parallel application of the model.

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 activate model sets of responses in parallel.

9. Concluding remarks

The demands on an effective Vision system are tremendous, and require a large number of well integrated functionalities. A processing system must be able to implement a model structure, the complexity of which is directly related to the structural complexity of the problem under consideration in the external world. One of the central functionalities is an ability to learn the external world through interaction and exploration. This will require totally different information representations and computing structures from what is common today.

Acknowledgements

The author wants to acknowledge the financial support of WITAS: The Wallenberg Laboratory For Information Technology and Autonomous Systems, as well as NUTEK: The Swedish National Board of Technical Development. These organizations 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 Linkoeping University, for discussion of the contents as well as for text and figure contributions to parts of the manuscript.

References

[1] D.H. Ballard, Animate vision, Technical Report 329, Computer Science Department, University of Rochester, February 1990.
[2] D.H. Ballard, C.M. Brown, Computer Vision, Prentice-Hall, Englewood Cliffs, NJ, 1982.
[3] D. Beymer, T. Poggio, Image representations for visual learning, Science 272 (June 1996) 1905–1909.
[4] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley-Interscience, New York, 1973.
[5] D. Gabor, Theory of communication, J. Inst. Electron. Eng. 93 (26) (1946) 429–457.
[6] M.S. Gazzaniga, The Social Brain, Discovering the Networks of the Mind, Basic Books, New York, 1985.
[7] G.H. Granlund, In search of a general picture processing operator, Comput. Graphics Image Process. 8 (2) (1978) 155–178.
[8] G.H. Granlund, Integrated analysis-response structures for robotics systems, Report LiTH-ISY-I-0932, Computer Vision Laboratory, Linköping University, Sweden, 1988.
[9] G.H. Granlund, H. Knutsson, Signal Processing for Computer Vision, Kluwer Academic Publishers, Dordrecht, 1995. ISBN 0-7923-9530-1.

[10] W.E.L. Grimson, Object Recognition by Computer: The Role of Geometric Constraints, MIT Press, Cambridge, MA, USA, 1990.
[11] L. Haglund, H. Knutsson, G.H. Granlund, Scale and orientation adaptive filtering, in: Proc. 8th Scandinavian Conf. on Image Analysis, Tromsö, Norway, May 1993. NOBIM, Report LiTH-ISY-I-1527, Linköping University.
[12] R.M. Haralick, L.G. Shapiro, Computer and Robot Vision, Vol. 2, Addison-Wesley, Reading, MA, 1993.
[13] R. Held, A. Hein, Movement-produced stimulation in the development of visually guided behavior, J. Comparative Physiol. Psychol. 56 (5) (October 1963) 872–876.
[14] B.K.P. Horn, Robot Vision, MIT Press, Cambridge, MA, 1986.
[15] D.H. Hubel, Eye, Brain and Vision, Scientific American Library, Vol. 22, Freeman and Company, San Francisco, CA, 1988. ISBN 0-7167-5020-1.
[16] D.H. Hubel, T.N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's striate cortex, J. Physiol. 160 (1962) 106–154.
[17] R.I.G. Hughes, The structure and interpretation of Quantum Mechanics, Harvard University Press, Cambridge, CA, 1989. ISBN: 0-674-84391-6.
[18] L. Jacobsson, H. Wechsler, A paradigm for invariant object recognition of brightness, optical flow and binocular disparity images, Pattern Recognition Letters 1 (October 1982) 61–68.
[19] K. Kanatani, Camera rotation invariance of image characteristics, Comput. Vision Graphics Image Proces. 39 (3) (September 1987) 328–354.
[20] L.C. Katz, C.J. Shatz, Synaptic activity and the construction of cortical circuits, Science 274 (15 November 1996) 1133–1138.
[21] J.J. Koenderink, 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 (1975) 773–791.
[22] J.J. Koenderink, A.J. van Doorn, The structure of images, Biol. Cybernet. 50 (1984) 363–370.
[23] T. Landelius, Behavior representation by growing a learning tree, Thesis No. 397, September 1993. ISBN 91-7871-166-5.
[24] T. Landelius, 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.
[25] T.Landelius, H. Knutsson, Behaviorism and reinforcement learning, 2nd Swedish Conf. on Connectionism, Skövde, March 1995, pp. 259–270.
[26] T. Landelius, 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.
[27] R.A. Lewitt, Physiological Psychology, Holt, Rinehart & Winston, New York, 1981.
[28] L.M. Lifshitz, Image segmentation via multiresolution extrema following, Tech. Report 87-012, University of North Carolina, 1987.
[29] R. Linsker, Development of feature-analyzing cells and their columnar organization in a layered self-adaptive network, in: R.M.L. Cotteril (Ed.), Computer Simulation in Brain Science, Chapter 27, Cambridge, Cambridge University Press, 1988, pp. 416–431.
[30] R. Milanese, Focus of attention in human vision: a survey, Technical Report 90.03, Computing Science Center, University of Geneva, Geneva, August 1990.
[31] M. Mishkin, L.G. Ungerleider, K.A. Macko, Object vision and spatial vision: Two cortical pathways, Trends Neurosci. 6 (1983) 414–417.
[32] J.L. Mundy, A. Zisserman (Eds.), Geometric Invariance in Computer Vision, MIT Press, Cambridge, MA, USA, 1992. ISBN 0-262-13285-0.
[33] K. Nordberg, G. Granlund, H. Knutsson, Representation and learning of invariance, in: Proc. IEEE Internat. Conf. on Image Processing, Austin, TX, IEEE, New York, November 1994.
[34] T. Poggio, S. Edelman, A network that learns to recognize three-dimensional objects, Nature 343 (1990) 263–266.
[35] J.L. Raymond, S.G. Lisberger, M.D. Mauk, The cerebellum: A neuronal learning machine?, Science 272 (May 1996) 1126–1131.
[36] G.M. Shepherd, The Synaptic Organization of the Brain, Second ed., Oxford University Press, Oxford, 1979.
[37] R.W. Sperry, Science and Moral Priority: Merging Mind, Brain and Human Values, Praeger, New York, 1985.
[38] S.P. Springer, G. Deutsch, Left Brain, Right Brain, Freeman, New York, 1993.
[39] S. Ullman, R. Basri, Recognition by linear combinations of models, IEEE Trans. Pattern Anal. Mach. Intel. 13 (10) (1991) 992–1006.
[40] C-J. Westelius. H. Knutsson, G.H. Granlund, Focus of attention control, in: Proc. 7th Scandinavian Conf. on Image Analysis, Aalborg, Denmark, August 1991, pp. 667–674, Pattern Recognition Society of Denmark.
[41] R. Wilson, G.H. Granlund, The uncertainty principle in image processing, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-6 (6) (November 1984). Report LiTH-ISY-I-0576, Computer Vision Laboratory, Linköping University, Sweden, 1983.
[42] A. Witkin, Scale-space filtering, in: Proc. 8th Internat. Joint Conf. Artificial Intelligence, Karlsruhe, 1983, pp. 1019–1022.