Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

A Cortical-type Modular Neural Network for Hypothetical Reasoning

Edgar Körner, Hiroshi Tsujino, Tomohiko Masutani

1997

Preprint:

This is an accepted article published in Neural Networks. The final authenticated version is available online at: https://doi.org/[DOI not available]

Pergamon

Neural Networks, Vol. 10, No. 5, pp. 791–814, 1997
© 1997 Elsevier Science Ltd. All rights reserved
Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(96)00126-8

CONTRIBUTED ARTICLE

A Cortical-type Modular Neural Network for Hypothetical Reasoning

EDGAR KOERNER, HIROSHI TSUJINO AND TOMOHIKO MASUTANI

Wako Research Center, HONDA R&D Co. Ltd

(*Received* 20 *November* 1995; *accepted* 25 *November* 1996)

Abstract—*We propose a multilayer neural network architecture that can implement the kind of hypothetical reasoning that the cortex seems to perform in making sense of the sensory input. The elementary processing nodes of each homogeneous sheet are not single formal neurons, but complex modules abducted from the functional organization of neocortical columns. As an example, we simulate face recognition in this neocortical architecture. A holistic but coarse initial hypothesis is generated by express forward input description and subsequently refined under the constraints of this hypothesis. Separation of forward input description and feedback generated hypothesis, while using the difference in both descriptions at each of the modular units to control the refinement, enables robust recognition and has the potential for autonomous learning. © 1997 Elsevier Science Ltd.*

Keywords—Modular neural network; Recurrent network; Visual cortex; Analysis-by-synthesis; Hypothetical reasoning; Signal–symbol processing; Object recognition.

1. SENSORY INTERPRETATION IS A PROCESS OF HYPOTHESES GENERATION AND VERIFICATION

1.1. Hypotheses constitute dynamic constraints for top-down control of sensory interpretation

The aim of our study is to produce an artificial neural system that can perform robust recognition and autonomous learning under real world conditions that regularly include complex and ambiguous scenes. Recognition of one out of a limited set of expected objects is well handled by parallel matching. Especially for face recognition problems it has been shown that a forward processing architecture performs sufficiently well (Brunelli & Poggio, 1993; Wiskott & von der Malsburg, 1995; Pentland et al., 1994). However, under real world conditions, a parallel recall may frequently face the problem of combinatorial explosion of possible alternatives. Triggering a partial match by local filtering does not always resolve such a decision impasse either, as is demonstrated with ambiguous figures like the famous "fuzzy dog". Therefore, recognition of complex sensory situations must be an active process involving both bottom-up and top-down processing.

The first stage in visual recognition seems to be a coarse holistic classification of the input (Rosch et al., 1976), which is subsequently identified in detail (Cavanagh, 1991). Experimental evidence in psychophysics supports the assumption that conscious experience involves an active modeling [for a review see Picton and Stuss (1994)]. The visual illusion of backward masking (metacontrast) provides compelling evidence that perception is not instantaneous and that it occurs sequentially in distinct stages that can be strongly modulated by top-down influences, although this phenomenon must have its roots early in visual processing (Ramachandran & Cobb, 1995). The top-down process may add knowledge to stimulus signals to enable the interpretation of the ghostly patterns of stimuli as solid objects in the external world (Gregory, 1995). Therefore, to perceive the world by means of analysis-by-synthesis, building an internal model of the external world and fitting this to the sensory information we receive seems to be a general characteristic of cortical processing in both its conscious and unconscious parts. Imagery and perception activate similar cortical areas in vision and audition (Kosslyn et al., 1993, 1995; Smith et al.,

Acknowledgements: The authors are grateful to Dr K. Tanaka, Laboratory of Neural Information Processing RIKEN, and two anonymous reviewers for critical comments and suggestions that considerably helped to improve the quality of the paper.

Requests for reprints should be sent to Prof. Dr-Ing. Dr Sci. Edgar Koerner, Chief Scientist, HONDA R&D Co. Ltd, Wako Research Center, 1-4-1 Chuo, Wako-shi, Saitama, 351-01, Japan; Phone: +81 48 461 2511; Fax: +81 48 462 5062; e-mail: edgar.koerner@f.rd.honda.co.jp.

1995) and motor images also have the same properties as the corresponding motor representations (Jeannerod, 1995).

Finding structure and regularity in the flow of sensory events, and giving meaning to the sensory input is the part the neocortex performs better than other brain structures. Aspects already understood should be removed from the data stream to make it easier to detect new regularities (Barlow, 1994; Koerner, 1994). For this purpose, the neocortex has to construct a working model of the environment to compare it with the sensory input, remove those parts that match known items and analyse the residue. Therefore, interpretation of sensory input seems to be not only the activation of a higher-level internal representation, but a kind of generation and verification of internal hypotheses on the input: a process of recreation of the internal representational architecture which relates best to the configuration of sensory input and which serves as emerging dynamic constraints for the further refinement of recognition.

The creation of an internal hypothesis and using it as a prediction of the next input reduces the high dimensional decision problem of the interpretation of the massively parallel sensory input to a simple comparison between two configurations: input and prediction. Otherwise it is impossible to behave as fast as living beings with cortex do.

1.2. Required characteristics of an artificial neural system capable of hypothetical reasoning

To perform the hypothetical reasoning of sensory input efficiently, the system has to have the following characteristics.

1. The feedforward flow of signals along the hierarchical system carries an input description, while the feedback flow represents an internally generated hypothesis. For the system to work from the beginning in the mode of comparison between the input and prediction, the system has to generate an initial hypothesis. The initial hypothesis must be a globally consistent description of the input, to limit the breadth of the search for finding proper hypotheses at any level of the hierarchy by prediction. Therefore, we propose that **an initial hypothesis emerges at the highest level of the hierarchy** whether the process is triggered by a sensory input as in recognition or from within the system as in imagery (Section 2.1.1).
2. Composed hypothesis and input description differ qualitatively; the emerging hypothesis is based on the limited set of already acquired knowledge, while the input description must have the potential to describe any possible input. We conclude that in an artificial neural system capable of hypothetical reasoning, from the lowest level up to the highest level of processing **a separate representation of input description and hypothesis** has to exist (Section 2.1.2).

   To generate and maintain an internal image (hypothesis on the input), there is no time left to search sequentially for the elementary processing nodes in the system that caused the difference defined by a global comparison. The definition of difference at individual nodes at each hierarchic level is required to modify the hypothesis at the pace of input changes. This is also a crucial prerequisite for autonomous learning. The hypothesis is generated by the already acquired knowledge distributed over the hierarchical structure of the system, and, therefore, the difference between the hypothesis and input description at individual nodes at each hierarchical level of the system is the required control signal to define what is new, and where it has to be learned.
3. High sensitivity to changes in the sensory input while maintaining a hypothesis is crucial for a timely decision on updating the current hypothesis. We propose **"switching-off" the part of the input description already interpreted** by the emerging hypothesis (Section 2.1.3). This is a reasonable assumption, since ongoing activation of the input description would mask emerging changes in the sensory environment. As is well known, this is not the case in the brain. Switching-off the interpreted part of the input description serves best to verify where the hypothesis has to be updated: the remaining input description supplies the energy to force the modification or the learning until the difference becomes small enough.

   We propose a modular artificial neural system composed of **elementary processing nodes corresponding to neocortical columns** (Section 2.2) which provides these characteristics.

1.3. Implementation of cognitive processes like hypothetical reasoning in neural architecture

In this paper we propose a neocortical type artificial neural system that serves to implement the above outlined characteristics (Section 2). The transition from local topographic filtering to object definition in visual object recognition is a proper model setup to investigate those characteristics, since (i) image recognition provides a proper scenario for interpretation of massively parallel, highly variable real world signal input; and (ii) the ventral visual pathway, which is responsible for object recognition, is one of the best investigated cortical systems providing experimental data to design and verify the model.

The internal image has to be generated to a resolution level required for the respective behavioral purpose. For most tasks it will be sufficient to establish the internal image and keep it as a prediction for on-line evaluation of the sensory input at a quite coarse resolution level

defining a situation or a general object class. However, for inspection and verification of an object's details it must be possible to establish the prediction downward to the highest possible resolution at the first cortical sensory areas. Therefore, the hypothetical reasoning should be expected to be a general characteristic of all hierarchical levels of neocortical processing in the system of object recognition — an assumption that is supported by the fact that the neocortical hardware (the columnar units and its basic internal organization) is rather similar at any level. This hypothesis is additionally supported by recent experiments on binocular rivalry: despite a rather constant input, the representation activated by dissimilar images submitted concurrently to the right and left eye seems to be switched in accordance with the percept at any level of the processing hierarchy down to the first cortical level (V1) (Leopold & Logothetis, 1996).

It has been stressed in several model approaches that bi-directional communication is an essential characteristic of neural processing. Recurrent connectivity between different hierarchic levels of recognition has been proposed as a means for relaxation-type refinement (Poggio et al., 1985; Mumford, 1992), attention control (Fukushima, 1986) or to control the recognition process for further refinement (Kawato et al., 1993; Ando, 1996; Rao & Ballard, 1995). The separation, however, between input description and hypothesis has not been addressed. Furthermore, in these previous models the subtraction of the feedback-generated image from the input image gave only the Euclidean distance in the pixel space, but did not define precisely where the hypothesis has to be updated — and by what part of the existing knowledge. Closest to our approach, at least with regard to image recognition, is the recently proposed *counter stream* model (Ullman, 1995) that proposes a separation of bottom-up and top-down processing streams, and a certain control of the course of the subsequent processing by "priming" the bottom-up processing stream by previous top-down passage at the same local unit, and vice versa. However, the type of processing performed by each node of the network and the representation of information in the network were not elaborate. We emphasize the control of internal communication that is important for evaluating multiple hypotheses, and the basic architecture that can resolve inconsistencies between input description and hypotheses.

Poggio and Hurlbert (1994) provided an elaborate discussion of forward processing in visual recognition, proposing that the system could be decomposed into elementary modules. Those modules were defined in terms of encoded knowledge, and in that respect we follow this general scheme of encoding. Conversely, we define the elementary node of our neural architecture as a module in terms of the local control mechanisms. A cluster of our modules may be involved in forming a modular unit of Poggio and Hurlbert (1994), but not exclusively. Furthermore, we do not limit iterative processing to higher representation levels. Our proposed modular artificial neural system is intended as a general scheme for reasoning in neural architecture.

In our implementation of cognitive-quality processing in artificial neural systems we follow suggestions by Minsky (1985) and Smolenski (1988). Parallel implementation of symbol processing in neural networks has previously been proposed (Hinton, 1991; Shastri & Ajjanagadde, 1993). However, the crucial problem in this paper is how to effect reasoning by neural architecture without just implementing conventional symbol processing (Barnden, 1995). We propose a "neocortical type" of formal neural architecture that performs hypothetical reasoning as its basic regime. There is a smooth transition between signal and symbol level processing. With signal we understand the analog activation pattern type of description of a feature, while the term symbol defines a set of patterns describing a class of features. In this paper we describe the architecture and the control mechanisms for hypothetical reasoning in our proposed neocortical architecture. Knowledge representation and its self-organization are dealt with in a subsequent paper (Koerner et al., submitted).

2. A NEOCORTICAL NEURAL NETWORK ARCHITECTURE FOR HYPOTHETICAL REASONING

The model proposed here does not attempt to account for a detailed description of a specific cortical structure, but we try to capture some general properties of architecture and processing that are relevant to hypothetical reasoning and object recognition in neocortical architecture. The problem to be solved is how internal models can be composed and verified by comparison to the sensory input, and what implementation in neural architecture is supported by both computational theory and experimental data.

2.1. Basic design philosophy: abducting some general features of neocortical architecture and processing

2.1.1. *Emergence of an initial hypothesis at the highest stage.* It is easily agreed that a proper global hypothesis on the sensory input will speed-up an image interpretation process and avoid the problem of combinatorial explosion that is inherent to the bottom-up local filtering approach. However, to find the proper global hypothesis is just the problem that the system has to solve.

According to the experimental evidence, we hypothesize that the brain avoids this dilemma by a coarse input to higher levels of the ventral visual pathway via the amygdala, triggering a crude holistic decision on the behavioral meaning of the sensory input (Fig. 1). Neurons in amygdala show visual stimulus-selective responses with a strong tendency to a particular category of stimuli of special behavioral meaning; such as the

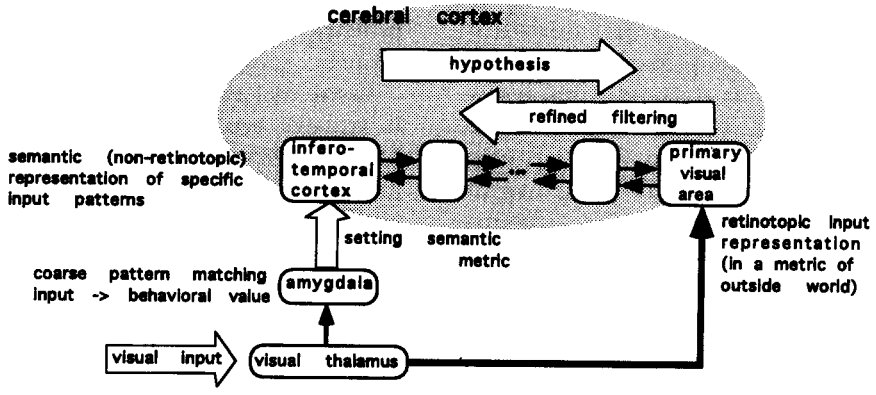

FIGURE 1. Activation of a coarse initial hypothesis by holistic evaluation of the input. The shortcut via the amygdala contacts the highest level of unimodal sensory processing, the inferotemporal cortex (IT), before (both phylogenetically and ontogenetically) the elaborate filtering via the cortical hierarchy makes contact to IT (Hernandez-Gonzales et al., 1994; Romanski & LeDoux, 1992). Representations in IT are defined in infant monkeys before the neocortical bottom-up refined sensory filtering gets access to the IT in ontogenetic development (Rodman, 1994). Amygdala is part of the value-system of the brain and known to code patterns of a behavioral specific value. Based on this experimental evidence, we propose that the shortcut input of raw visual data to the highest level of visual representation activates a coarse hypothesis which feeds forwards an expectation to the lower levels that sets the dynamic constraints for the refinement of this holistic but coarse hypothesis.

smiling face of a human, a particular person, a dry food, a threatening face of a monkey, etc. (Mikami et al., 1994; Adolphs et al., 1994). The rapid activation of a holistic initial hypothesis on the raw input data at the top level can impose content-dependent constraints on the hypothesis description of the lower levels of the neocortical hierarchy by facilitating the input description at selected (expected) locations. We assume that the refinement of the hypothesis description of the lower level is only possible if the respective higher level has a sufficiently entrained hypothesis to feedback a definite support to a local site to enable the forward activation of related details, limiting the breadth of the search and preventing combinatorial explosion of possible alternatives in local filtering.

This assumption is supported by the fact that object representations in the inferotemporal cortex (IT) are activated within 5 ms after arrival of the afferent information and do not change (Oram & Perrett, 1992), while at V1 and extrastriate visual areas the response of feature detectors may undergo a modulation (Dinse, 1994; Shevelev et al., 1993; but there is also Celebrini et al., 1993) and that for a majority of V1 cells the spatial structure of the receptive field changes as a function of time within 200 ms after the onset of the response (Freeman, 1995). We propose that the later activation is triggered by less salient parts of the input description under guidance from the feedback from the more global evaluation at the higher level, and that the 200 ms response duration reflects the refinement of the internal hypothesis.

2.1.2. *Separation of input description and hypothesis.* To implement the characteristics required for hypothetical reasoning in neural architecture (as described in Section 1.2), a replicator-like doubling of networks (Ullman, 1995; Hecht-Nielssen, 1995) with identical structures for bottom-up and top-down processing is necessary but not sufficient. The respective pairs of nodes must represent different qualities of description of the same content. Then, any of those pairs of nodes has to decide locally — according to comparison of its state with the state of the global system provided by the feedforward and feedback input — which of its possible states it should instantiate, and which of its representations has to be updated in the process of learning. The elementary processing node of networks capable of hypothetical reasoning must have a sufficiently complex organization both for local control and knowledge representation to act as the required "local agent", and this basic organization should be roughly the same across all nodes of the system regardless of its position in the hierarchy.

We propose the neocortical columnar module to be a model for the required complex elementary processing node. Our system architecture consists of a hierarchy of homogeneous networks composed of the neocortical columnar modules as the elementary processing nodes. Functionally, the system is composed of two parts that are locally connected at individual modular units. The hypotheses are generated from afferent input description in a forward processing hierarchy, and they are verified and updated in a recurrently connected system part that serves to recreate the hierarchy of dynamically linked modular units (Fig. 2).

2.1.3. *Switching-off the already interpreted part of the input.* Verification of the hypothesis is performed at each modular unit of the network system by the feedback prediction: the nodes (modular units) that have a forward activated hypothesis consistent with the predictive

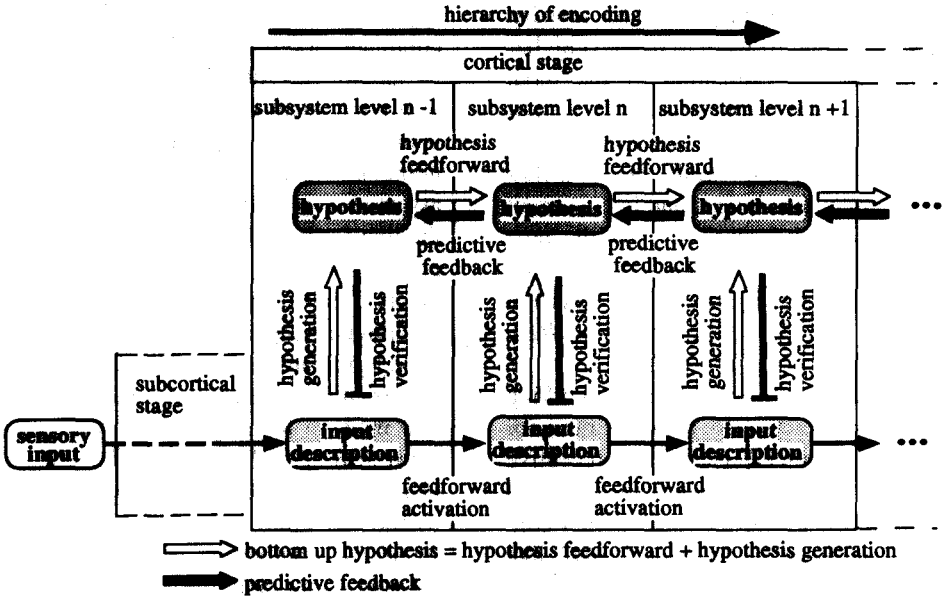

FIGURE 2. Simplified scheme of processing characteristics in a general neocortical-type artificial neural system. The subsystems at any hierarchic level are composed of modular units which have the same general local control for intra- and inter-modular communication. The modular units are abducted from the cortical architecture as the elementary computational units of neocortical type processing. The qualitative difference between subcortical and neocortical types of processing is roughly sketched.

feedback can strongly entrain this hypothesis, while switching-off the afferent input descriptions that generated this hypothesis. The local hypothesis is kept alive by mutual excitation of all local hypotheses that form a consistent global one within the recurrently connected hypothesis system part (Fig. 2).

2.2. A model neocortical network architecture

2.2.1. *Columnar modules as elementary processing nodes in neocortical-type neural networks.* Flattening the cortex we find an almost homogeneous sheet composed of complex nodes, columnar units, the basic internal organization of which does not depend on the information represented there. We propose that this unitary architecture of columnar units represents not the structure of the knowledge stored there, but the control that forces the system to make that representation.

We formalize a columnar unit of nth encoding level into three functional different submodules: $R0_{(n)}$ — coarse input evaluation of whether or not to allow the afferent input to activate $R1_{(n)}$; $R1_{(n)}$ — hypothesis

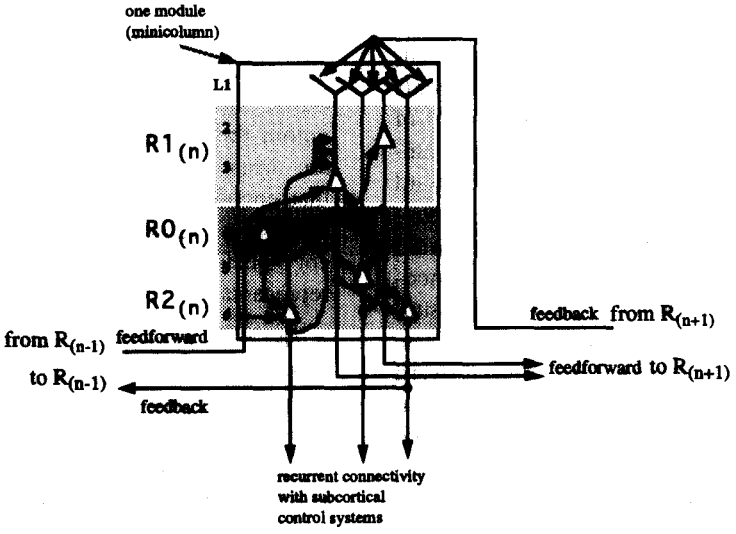

FIGURE 3. Schematic description of the architecture of elementary nodes in the neocortical-type artificial neural system, and its functional connectivity. The submodules within the elementary modules are abducted from general features of the neocortical columns: *R0* — layer 4; *R1* — supragranular layers 3 and 2; *R2* — layer 5 and 6.

description by compact, detailed encoding of the afferent input description, which is fed forward to the next higher level of encoding; $R2_{(n)}$ — evaluation of the input description and the predictive feedback from $R2_{(n+1)}$ of the next higher level of encoding. $R2_{(n)}$ is activated if the activation pattern of working hypothesis $R1_{(n)}$ matches quite sufficiently the input description, and is consistent with the predictive feedback from $R2_{(n+1)}$ of the next higher level. Then $R2_{(n)}$ switches-off $R0_{(n)}$, while being itself kept active at a lower activity rate by the recurrent connectivity within the $R2_{(n+1)} - R2_{(n)}$, $R1_{(n+1)} - R1_{(n)}$ subsystems (see Fig. 3).

$R2_{(n)}$ cannot have detailed information on the working hypothesis. $R2_{(n)}$ is always activated if one feature (or object) encoded by this column is present in the receptive field of the column, and if this decision is also confirmed by the predictive feedback within the more global context. Therefore, we assume $R2_{(n)}$ to be a symbol like representation of the feature encoded by the column, which can take many different instantiations that are defined by the respective details in the signal quality description at $R1_{(n)}$. As a consequence, the feedforward by $R1_{(n)}$ to $R0_{(n+1)}$ is of signal quality, always reflecting the current input situation in any detailed variation, while the feedback prediction via $R2_{(n)}$ to the next lower level modules is of symbol quality, connecting to the modules which have to be active if the feature encoded in its module is true — but the specific instantiation of those subsymbols is described by the working hypothesis $R1_{(n-1)}$ according to the current afferent input. Any possible instantiation of the subsymbols is supported and fixed as the current valid instantiation, while the detailed description of this instantiation at $R1_{(n)}$ is used in making up the next level's forward signal description at $R0_{(n+1)}$ and $R1_{(n+1)}$.

It is a plausible assumption that learning includes both the specific encoding of forward descriptions and the formation of feedback connections to the modules involved in the forward descriptions. After the representation is sufficiently established, switching-off prevents overlearning simply by lowering the activation level of the sustained hypothesis in the $R1 \leftrightarrow R2$ hypothesis subsystem of the architecture. As long as both feedforward and feedback connections are not yet established, neurons in lower level modules that provide the feedforward descriptions, and the neurons from the higher level module (the winner module) that provide the feedback support, continue to be active at the same time and can cause learning according to a Hebbian rule. The forward description of afferent information must have the potential to describe even situations not yet seen before — otherwise no learning and acquisition of novel information would be possible. Hence, $R1$ must be an objective description that satisfies this condition. The principal component analysis (PCA) scheme is definitely of that type, and we use it in our simulation (Section 3) as a convenient scheme for knowledge representation in $R1$.

Figure 3 shows the correspondence of the proposed model columnar module as an elementary processing node with its submodules $R0$, $R1$ and $R2$ to the general architecture of a neocortical column. Afferent input is evaluated by neurons in layer 4 ($R0$) that trigger activation of supragranular neurons ($R1$) of the same column, which are assumed to provide a detailed hypothesis description of the input using additionally direct afferent input. A subpopulation of pyramidal neurons in the infragranular layers 5 and 6 extends its apical dendrites through the column until layer 1, where the feedback from the higher level terminates massively (but also in layers 6, 3 and 2). Furthermore, all pyramids of layers 3 and 2 give collateral input to lower pyramids when passing its feedforward to $R0_{(n+1)}$. We assume that $R2_{(n)}$ (the pyramids in layer 5 and 6) evaluates the forward description provided by the activation state of $R1_{(n)}$ (pyramids in layer 3 and 2). If $R1_{(n)}$ activation pattern sufficiently matches the afferent input and the predictive feedback from the next level, $R2_{(n)}$ is maximally entrained and may switch-off $R0_{(n)}$ eliminating the driving input to $R1_{(n)}$ as we assume, possibly via recurrent collaterals to inhibitory neurons in layer 4 [feedback activation causes depression of the activity in the supragranular layers (Allison et al., 1995; Rockland & VanHoesen, 1994)]. The feedback to layer 1 of its respective lower level modules contacts the apical dendrites of pyramidal neurons of layers 2, 3, 5, 6 — but not to neurons of layer 4 and those pyramids of layer 6 which get direct afferent input (Cauller & Connors, 1994). Although supragranular layers contribute the majority of forward projections to the higher level (90%), a minor share is also from infragranular layers (10%), whereas this is opposite in the case of backward projection (90% deep layers and 10% upper layers) (Felleman & VanEssen, 1991). We focus here only on the dominant projections for reasons of simplicity.

2.2.2. *Basic control mechanisms in the modular artificial neural system.* Figure 4 illustrates the five types of interaction between elementary computational modules within the hierarchy of the system. The modular units are arranged into functional clusters, macro-columns, that encode different local features on the same local site of the "retinal" topography of the input, and different objects of the same class in the later non-topographic representation levels.

- (A) *Excitatory local control.* Within a local neighborhood, modules that code the same class of features, but do not get the same input vector (different receptive fields) support each other. As a result, the generation of a hypothesis at the modules is facilitated.
- (B) *Inhibitory local control.* Modules within a macro-column are mutually competitive. This serves to resolve local inconsistency; the competition between alternative features within the same macro-column is enhanced to a winner-takes-all characteristic in the

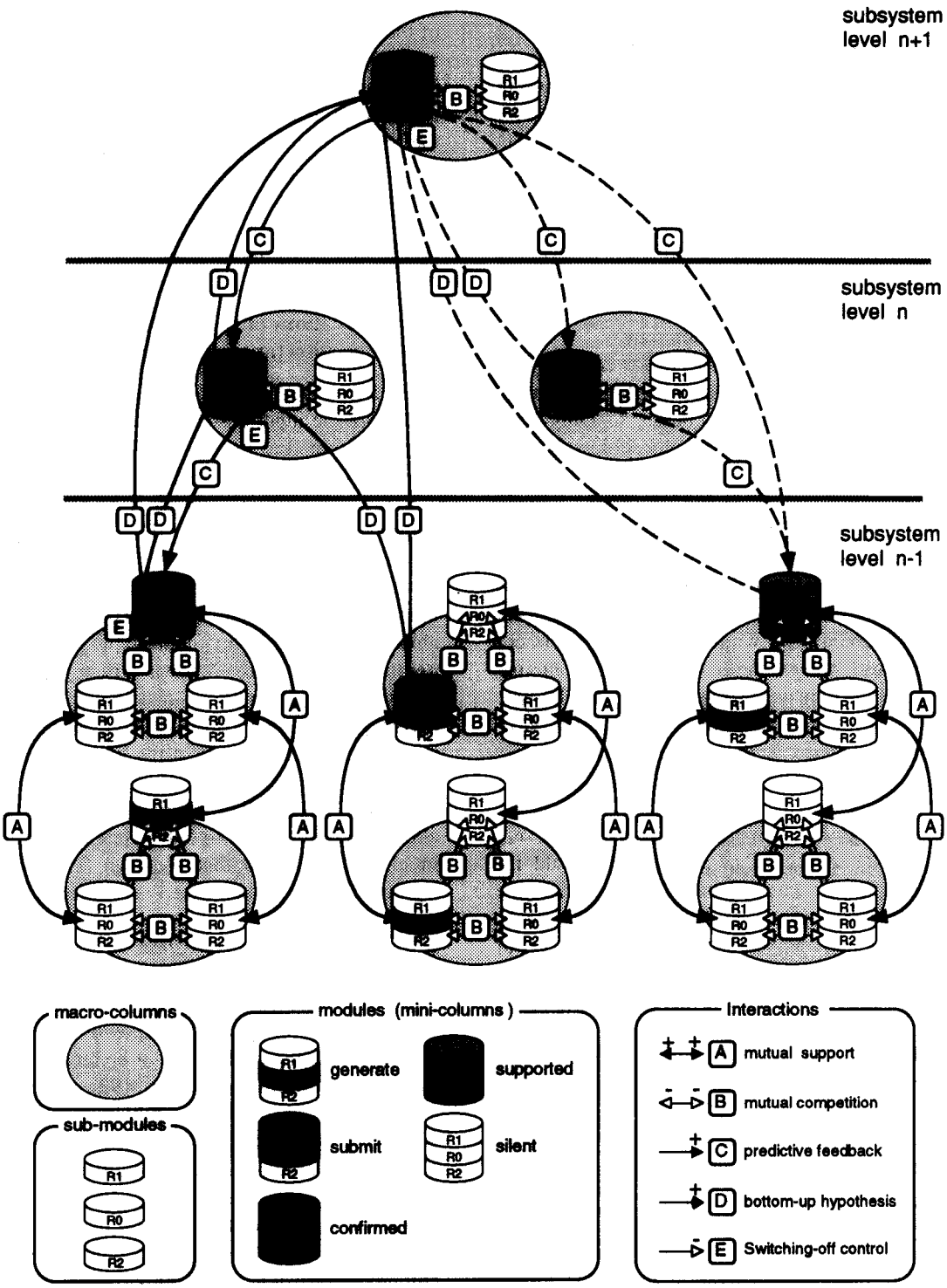

FIGURE 4. The interaction between elementary computational modules. Each module consists of *R0*, *R1* and *R2*. The modules corresponding to the neocortical mini-column constitute a cluster depicted by shadowed unit corresponding to the macro-column. The layers divided by the horizontal lines represent the subsystems which correspond to the level of encoding. The black and white arrows represent the excitatory and inhibitory controls between modules, respectively. The state of modules is changed by these controls. An active *R0* inhibits *R0*s of other modules in the same macro-column by control (B) to suppress alternative hypotheses, and supports *R0*s of other modules of the same class in next-neighbor macro-columns by control (A). Only a module with active *R0* can generate a hypothesis at *R1* and submit the hypothesis to modules in higher level subsystem by control (D). If the working hypothesis representation is successfully verified by *R2* of the same module, and is confirmed by the predictive feedback control (C) from modules in higher level subsystem, *R2* of the module starts to inhibit *R0* of the same module by control (E) and keeps itself active with lower activity. Then, one *R0* of other modules may become active by decreasing inhibition control (B) from *R0* of the already confirmed module. If the newly activated module is confirmed more strongly than the previous one, the previously confirmed module is inhibited by both *R2* of the new confirmed one by control (B) and the modules in higher level subsystem by control (C).

process of hypothesis verification by top-down and bottom-up control.

- (C) *Top-down control.* Active modules support those modules of the respective next lower-level subsystem that constitute local hypotheses consistent with their hypotheses, and suppress those that submit inconsistent local hypotheses by predictive feedback. If a hypothesis at a module is strongly activated, it can activate the respective lower-level modules by the predictive feedback even if the forward description there is fuzzy. The feature represented by a module in the higher-level subsystem activates all lower level descriptions that constitute its hierarchy.
- (D) *Bottom-up control.* Modules in lower-level subsystems forward their hypotheses to higher-level modules, the receptive field that they belong to. Only those modules activated above a certain threshold value can forward hypotheses.
- (E) *Switching-off control.* Switching-off the already interpreted part of the input description (see Section 2.1.3) has one additional consequence: we assume that switching-off the driving input description results in a markedly lower activation state of the respective local hypothesis. Accordingly, while strongly forward activated hypotheses prevent the emergence of any alternative description within its macro-column by the (B)-type inhibitory control, and also prevent the activation of alternative hypotheses at the neighboring local sites, verified hypotheses do not, and enable the refinement.

Additionally, a global normalization in the system (not shown in Fig. 4) serves as an adaptive threshold for hypothesis generation at all *R*0-submodules. In the case of complex or fuzzy inputs, several alternative hypotheses may emerge and compete at any hierarchic level of encoding. To prevent deadlock by combinatorial explosion of the (forward) hypothesis generation process, the number of activated alternative hypotheses should be limited, and the number of tolerated hypotheses should approach one with increasing activation value at *R*0 submodules. The inhibition at any *R*0 submodule is defined by the sum of overall activation at all *R*0 of the subsystem that means it is a locally applied global normalization. If activation of individual hypotheses is low, the inhibitory masking of *R*0 activation is also low, tolerating a quite large number of concurrently activated *R*0 submodules. With increasing activation level, the global normalization will silence more and more *R*0 that are less activated than *R*0 that may generate the consistent hypothesis. The dropping of less activated local hypotheses (*R*0) by increasing inhibitory masking is a syntactic evaluation of the reliability of local recognition of the input. The selection of the most reliable local hypothesis in the forward evaluation is increasingly biased by the emerging global hypothesis. (C)-type top-down control locally suppresses the inhibitory mask at those *R*0 submodules that are expected to be active if the global hypothesis is true, carving even weak *R*0 activation out of the inhibitory block [see eqn (1), Section 3.2.2]. This type of global normalization enables the activation of an initial hypothesis by reliable local hypotheses only, and the later refinement of the global hypothesis by selecting the proper ones out of the less reliable local hypotheses that otherwise could interfere with the convergence of the decision process. With this characteristic, the system performs a limited-breadth search while moving up (thinning-out) the decision tree.

2.3. Implications of the proposed artificial neural system architecture for autonomous recognition

2.3.1. *Robust recognition by recoding the sensory input.* The more the input description of a module fits the feature represented at this module, the earlier that module activates its hypothesis, and will have the best chance to win the competition. This is a reasonable model for neural recognition processes. The complexity of the sensory input is reduced by starting first with that part of the input that can be most reliably recognized to generate an initial (even if coarse) hypothesis. Switching-off the already interpreted part of the input description at any iteration removes the known structures of the input description to enhance the residual, maybe less salient parts for subsequent analysis. Then, the next rate salient local features can emerge at the lower-level subsystems — but under steady control of the global hypothesis being active at the *R*2 submodules. In this way the parallel input is decomposed into sequences of semi-parallel input ranked within that sequence according to their "probability of being correctly recognized".

2.3.2. *Autonomous acquisition of knowledge.* In our model, any input is decomposed into a set of already acquired patterns, creating and maintaining a consistent relational structure within the knowledge representation, and the residual part that cannot be reconstructed by the available knowledge (Koerner, 1994). The isolation of the residual part provides a highly specific teacher — which is a key problem for any autonomous learning. Even if one may argue that a simple priming between nodes of the afferent and efferent pathway (Ullman, 1995) could provide some control for routing and learning, those advanced learning mechanisms and control to enforce a consistent relational structure of the acquired knowledge of the system require a much more elaborate system architecture as the one proposed here.

3. HYPOTHETICAL REASONING IN A SIMPLIFIED PRIMATE VENTRAL VISUAL PATHWAY MODEL SYSTEM

To show the implementation of the modular architecture and its control proposed in Section 2, we simulated object recognition in this type of artificial neural

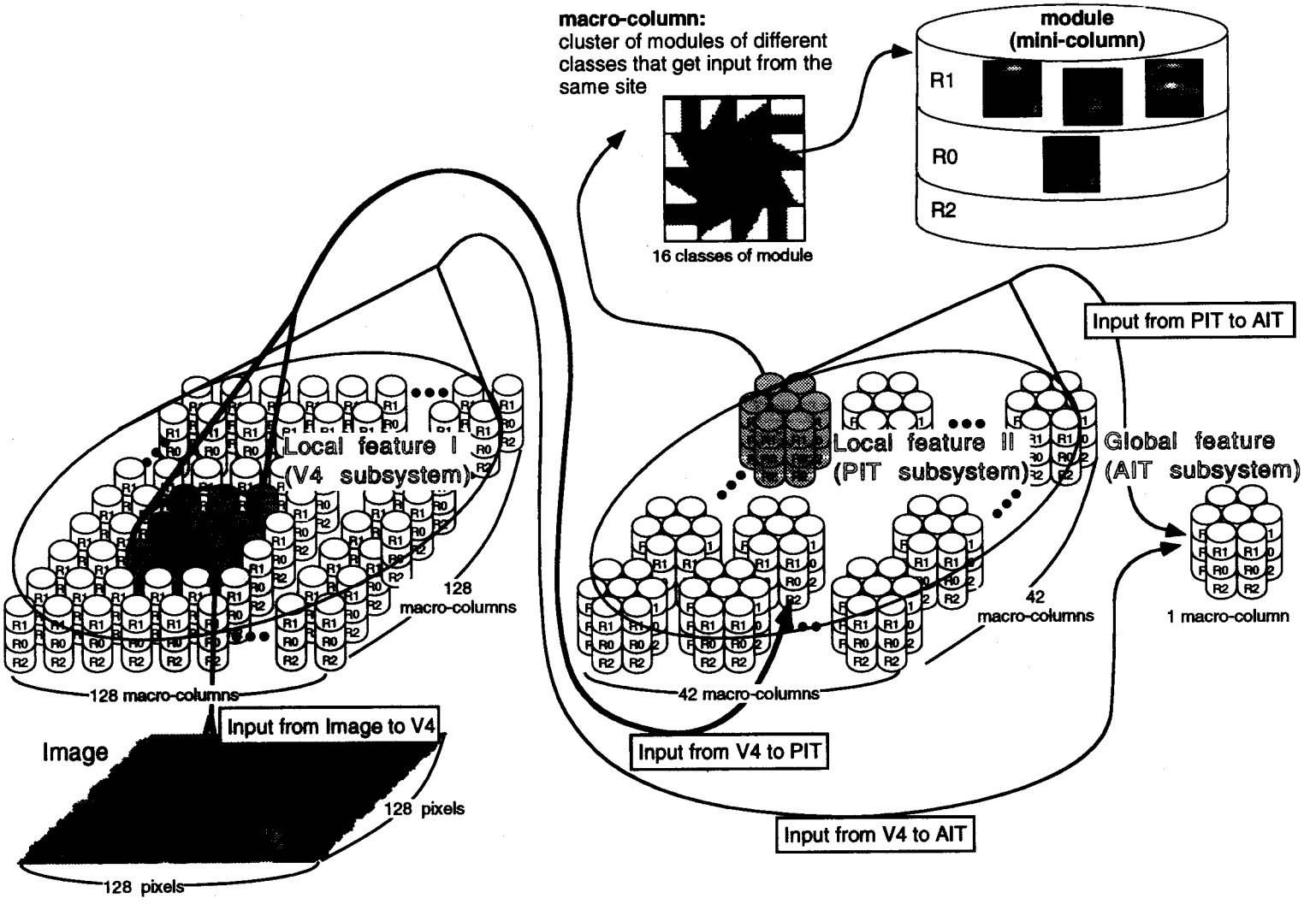

FIGURE 5. Characteristics of local and global encoding in the subsystems. The simulation system consists of the subsystems AIT, PIT and V4. The detailed configuration is described in Appendix A. The input image is shown bottom left of the figure. Each subsystem is defined by an homogeneous array of elementary modules represented by circles. A module corresponds to a mini-column, and a cluster of modules of different classes getting the same input description corresponds to a macro-column. Modules in V4, where the macro-column consists of only one module, represent an input image by "general-PCA" and submit the information to both PIT and AIT. In PIT, each macro-column consists of 16 modules that describe different classes of local features, as shown in the upper right figure.

system. We limit the implementation of the simulation to demonstrate the most important characteristics of our system, i.e. the top-down processing ability and the significance of the separation of input description and hypothesis. Because we limit the class of object to the face, we demonstrate the basic function of the proposed architecture, but we cannot show the explicit effect of the sophisticated control of the hypotheses.

3.1. Functional architecture of the system

The task is limited to a simple classification and identification of faces in an expected format. We focused on the definition of a global object from extracted local features which seems to take place at the transition from the posterior inferotemporal cortex (PIT) to the anterior inferotemporal cortex (AIT) in monkeys (Tanaka, 1993). The system consists of three subsystems: V4, PIT and AIT (Fig. 5). Because of limited computational resources, we neglected lower levels of local filtering and their functions are put together in one level of local feature extraction in V4. Any subsystem consists of an homogeneous array of modules described in Section 2. The respective size of the area at the input image processed by one module of each subsystem is different. Modules in V4 operate on the local patch of input, modules in PIT still perform a local topographic filtering. The receptive field of modules in AIT is assumed to cover all central vision relevant for object recognition to establish a non-topographic representation of global input features covering all modules of V4 and PIT. Both the effect of priming and coarse selection of a holistic decision via short-cut input by amygdala and "express-features" of the cortical local filtering are implemented to the direct image input from V4 to AIT subsystem. Next, we describe the function of each subsystem (see Appendix A for details).

3.1.1. *V4-subsystem: local feature I.* V4 is composed of modules representing the input description by some primitive feature, such as color, orientation, etc. We simulated a simplified V4 subsystem that serves (i) local filtering of the input image; and (ii) segmentation of the object by the predictive feedback.

3.1.2. *PIT-subsystem: local feature II.* PIT has local recognition modules that recognize the object in cooperation with AIT. Cells in PIT have more elaborate knowledge than those in V4 and smaller receptive fields than those in AIT (Kobatake & Tanaka, 1994; Tanaka et al., 1991). We consider the relation between AIT and PIT as a "frame" structure introduced by Minsky (1975); AIT defines global knowledge of the object, and local parts of the object are described in PIT. According to neurobiological evidence (Fujita & Fujita, 1996), we assume that "part-of" links between AIT and PIT are established. AIT also describes the spatial relation between the knowledge represented locally in PIT preserving the correspondence between the local features. PIT has an important role when we have to distinguish the object in the same class because it has more detailed localized knowledge than AIT.

3.1.3. *AIT-subsystem: global feature.* AIT modules have a large receptive field and the description at AIT shows some invariant properties (location, scale and rotation).

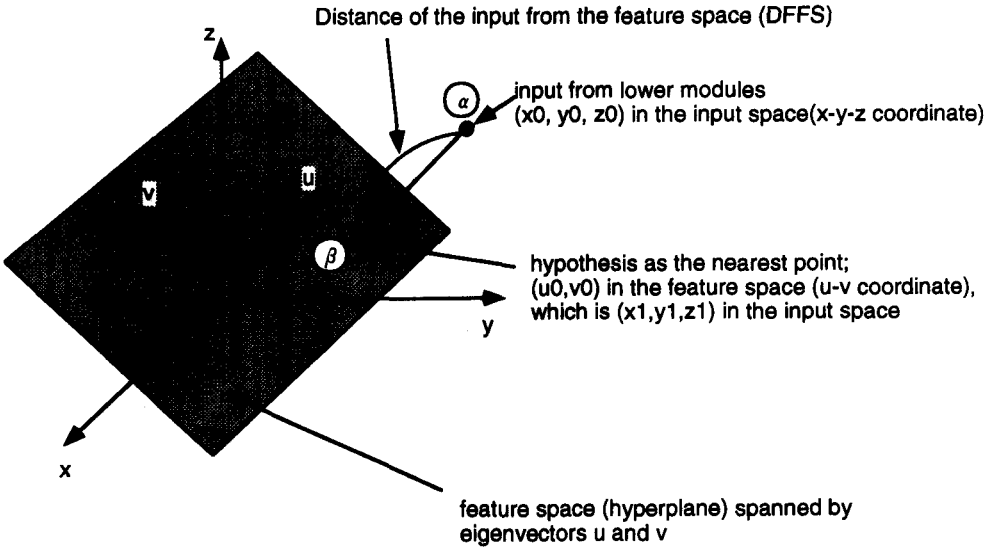

FIGURE 6. Eigenspace and hypothesis. *x–y–z* coordinate system represents the input space made up by converging output of lower-level modules. *u–v* coordinate system represents the feature space spanned at *R*1. (α) represents the input description coming from the lower-level modules. (β) represents the nearest hypothesis description which is the orthogonal projection of (α) onto the feature space spanned by the eigenvectors of *R*1.

Except for these differences, the basic structure of AIT is the same as the PIT subsystem. AIT gets input from both V4 and PIT. By the information from V4, AIT produces an initial coarse hypothesis about the object currently at the input. The resolution of the hypothesis is adaptively changed by the relation between the input and the state of the hypothesis. At the beginning, the profile of the reconstructed input vector is just a mean input vector, or mean face. As the process of iterative refinement of the hypothesis goes on, *R*1 is becoming active and the reconstructed input vector gets a finer resolution as the hypothesis on the input from PIT.

3.2. Elementary computational module

3.2.1. *Knowledge representation: eigenspace.* We apply the eigenspace method (Turk & Pentland, 1991) for the knowledge representation of the module. The eigenspace

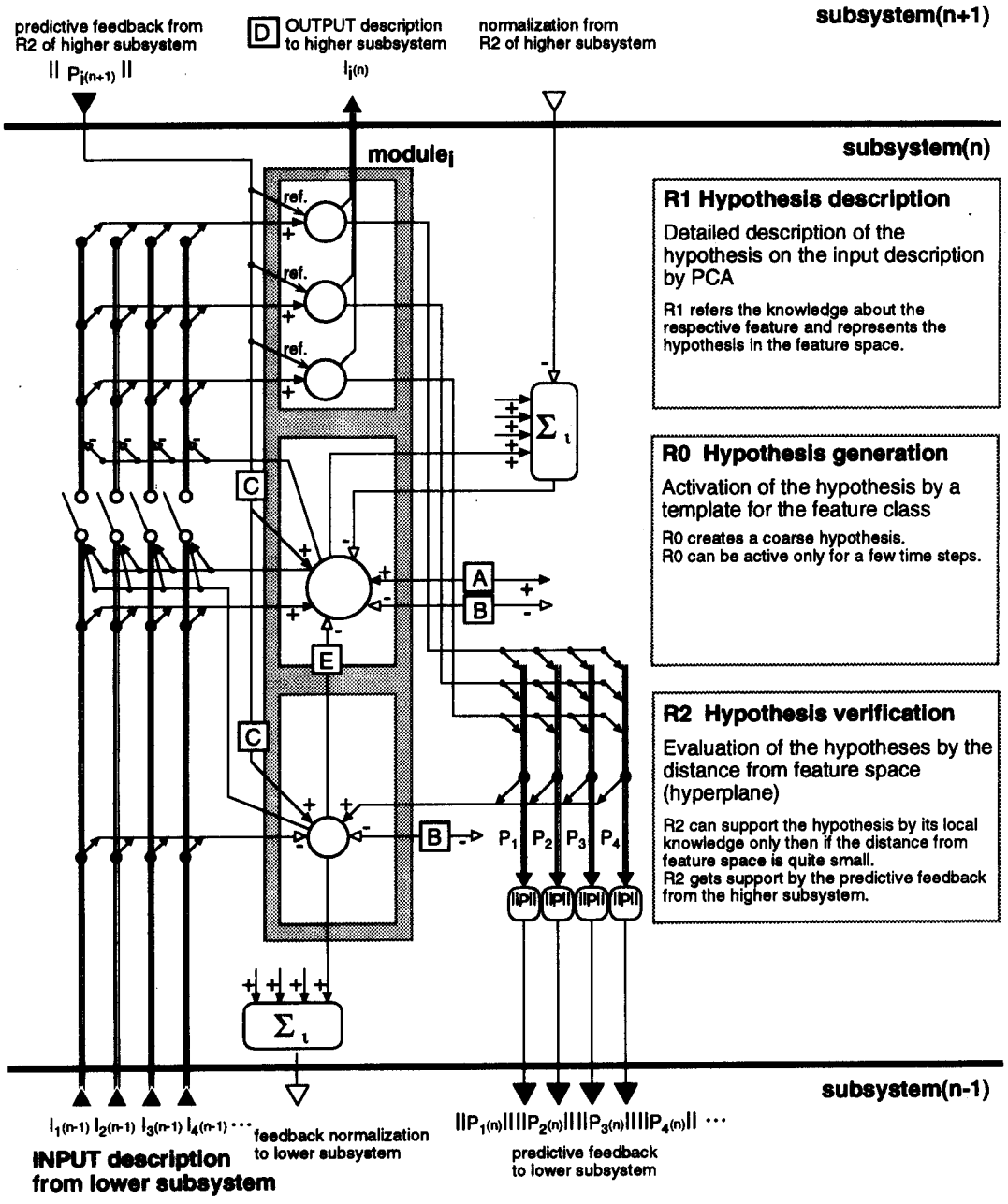

FIGURE 7. The elementary computational module. The elementary computational module consists of the subsystems *R*0, *R*1 and *R*2 as described in Section 2.2. The connectivity is shown for one module. Capital letters, (A)–(E), refer to control and signal flow as described in Section 2.2.2. The module gets its afferent input $I_{(n-1)}$ from the lower modules where each input $I_{j(n-1)}$ represented by triple-lines is a vector from *R*1s of the lower modules, *module*$_{j(n-1)}$. A predictive feedback, $\| P_{(n)} \|$, is prepared to serve a norm of vector $P_{j(n)}$ to the respective lower modules *j*. The hypothesis $I_{j(n)}$ is forwarded to the next higher level module from which the prediction $P_{j(n+1)}$ returns.

of a specific feature is computed by PCA on the training data of this feature. We call the space spanned by these eigenvectors the feature space.

Figure 6 illustrates how the hypotheses are treated on the feature space. (α) represents the input description coming from the lower-level modules. (β) represents the nearest hypothesis description in the feature space that is spanned by the eigenvectors of $R1$. At $R2$, the match of (α) against (β) is evaluated by the distance from feature space (DFFS) (Pentland et al., 1994). The generated hypothesis description (β) is used for both the bottom-up hypothesis to the higher-level modules as ($u0$, $v0$) in the feature space and the predictive feedback to the lower-level modules as ($x1$,$y1$,$z1$) in the input space.

3.2.2. *Submodules of the elementary computational module.* The elementary computational module (Fig. 7) consists of three submodules $R0$, $R1$ and $R2$. The simulated modules are simplified and computationally optimized with regard to the selected knowledge representation by PCA. Therefore, the module is functionally equivalent with the structure proposed in Section 2.2 and Figs 3 and 4, but nodes of the submodules do not directly correspond to neurons in the biological original. Moreover, according to PCA, the mean vector of each PCA is subtracted from the input to $R1$, and again added for the predictive feedback by $R2$, which is an artificial setting departing from the model in Section 2.2 and Figs 3 and 4. To simplify control, the predictive feedback from $R2_{(n+1)}$ to the lower level modules is provided by the norm of the reconstructed input vector (for further details, see (C)-type top-down control in Appendix B). The predictive feedback from $R2_{(n+1)}$ controls not only $R2_{(n)}$ but also $R0_{(n)}$, and is refered to by $R1_{(n)}$. In deviation from the scheme of Fig. 3, we added feedback also to $R0_{(n)}$ in our artificial neural system since it is very effective in supporting the hypothesis generation.

3.2.2.1. R0: hypothesis generation. Equation (1) describes $R0_{(n)}$ of the nth level of encoding which has a template, $template_{(n)}[k]$, to apply a kind of fuzzy classification on the input description by simple template matching. In this equation, the activation of $R0_{(n)}$ represents the analog strength of the related hypothesis and is controlled by the cooperating local hypotheses ((A)-type excitatory local control), the competing local hypotheses ((B)-type inhibitory local control), the predictive feedback from $R2_{(n+1)}$ of the respective module of the next higher subsystem ((C)-type top-down control), the feedback from $R2_{(n)}$ of the same module ((E)-type switching-off control) and global normalization. The effectness of (C)-type control to locally suppress the global normalization at those modules that are expected to belong to the global hypothesis (see Section 2.2.2) is higher for better respective $R1_{(n+1)}$ matches $I_{(n)}$.

The subscript n represents the level of the subsystem. The suffix k ($0 \le k < K$) represents the hypothesis class. y ($0 \le y < Y$) and x ($0 \le x < X$) represent the position of macro-columns in this subsystem. In the lower subsystem like V1, V2, V4 and PIT, the macro-column corresponds to the retinal position, but in AIT, the macro-column corresponds to the iconic pattern (Tanaka et al., 1991). Here t is the time variable. The hard-wired weights are set as $w_1 = 0.5$ and $w_2 = 0.2$ for all $R0$. See Appendix B for the detailed definition and explanation of controls in eqn (1).

$$
\begin{aligned}
& R0_{(n)}[k][y][x](t) \qquad (1)\\
&= f(\text{the value based on the distance between}\\
&\quad \text{the template } k \text{ and local input.}\\
&\quad + \text{(A)} - \text{type excitatory local control}\\
&\quad - \text{(B)} - \text{type inhibitory local control}\\
&\quad + \text{(C)} - \text{type top} - \text{down control from } R2 \text{ of the next}\\
&\quad \text{higher subsystem}\\
&\quad - \text{(E)} - \text{type switching} - \text{off control}\\
&\quad - \text{global normalization}\\
&= f(template_{(n)}[k] \cdot I_{(n-1)}[y][x](t-1)\\
&\quad + w_1 \times \sum_{\text{neighbor}yy,xx} R0_{(n)}[k][yy][xx](t-1)\\
&\quad - \max_{kk \neq k} R0_{(n)}[kk][y][x](t-1)\\
&\quad + w_2 \times \| P_{(n+1)}[k][y][x](t-1) \|\\
&\quad - R2_{(n)}[k][y][x](t-1)\\
&\quad - \max\Big(\sum_{\forall kk,yy,xx} R0_{(n)}[kk][yy][xx](t-1)\\
&\quad - \frac{\sum_{\forall l,u,v} R2_{(n+1)}[l][u][v](t-1)}{L}, 0\Big),
\end{aligned}
$$

where f is the monotone increasing function, $f \in [0,1]$,

"×" denotes a dot product operator,
$template_{(n)}[k]$ is the M-dimensional vector of the template,
$I_{(n-1)}[y][x]$ is the M-dimensional vector of the input description in the receptive field of $R0$ from the $(n-1)$th subsystem,
$P_{(n+1)}[k][y][x]$ is the J (see Section 3.2.2.2) dimensional vector of the predictive feedback from $R2$ of the $(n+1)$th subsystem,
$R2_{(n+1)}[l][u][v]$ is the scalar of $R2$ activity at the $(n+1)$th subsystem where $0 \le l < L$, $0 \le u < U$, $0 \le v < V$.

3.2.2.2. $R1$: hypothesis description. $R1$ analyzes and represents the hypothesis that is generated by related $R0$. Equation (2) describes the $R1$. The subscript n and the suffixes k, y, x and t are the same as those for $R0$, and j ($0 \leq j < J$) corresponds to the jth principal component represented by $PC_{(n)}[k][j]$.

$$R1_{(n)}[k][y][x][j](t) = \begin{cases} \text{projected information on the } j\text{th eigenvector} \\ \quad (\text{if } R0_{(n)}[k][y][x](t-1) \text{ is creating the hypothesis}) \\ \quad \text{stabilizing previous hypothesis} \\ (\text{if the hypothesis is confirmed by the predictive feedback and} R0_{(n)}[k][y][x](t-1) \\ \quad \text{is switched off.}) \end{cases}$$

$$= \begin{cases} PC_{(n)}[k][j] \cdot (I_{(n-1)}[y][x](t-1) - \text{Mean}_{(n)}[k]) \\ \quad (\text{if } R0_{(n)}[k][y][x](t-1) \geq R0_{\text{threshold}}) \\ \quad R1_{(n)}[k][y][x][j](t-1) \\ \quad (\text{if } R2_{(n)}[k][y][x](t-1 \geq 1.0 \text{ and} \\ \quad \|P_{(n+1)}[k][y][x]\| > 0), \end{cases} \quad (2)$$

where $PC_{(n)}[k][j]$ is the jth eigenvector of class k, and $\text{Mean}_{(n)}[k]$ is the M-dimensional mean vector of class k.

3.2.2.3. $R2$: hypothesis verification. $R2$ verifies the current hypothesis by comparing the input description and the working hypothesis. We use DFFS as a tool for evaluating the match between the input description and the working hypothesis (see Section 3.2.1 and Fig. 6). Equation (3) describes $R2$. In this equation, the distance between the $R1$ representation and the feature space is calculated by the second line. $R2$ gets the feedback from the next higher subsystem to enable more global or combinatorial verification.

The subscript n and the suffixes k, y, x and t are the same as those for $R0$.

$$\begin{aligned} & R2_{(n)}[k][y][x](t) \\ & = \text{normalization factor} \\ & - \text{DFFS} \\ & - \text{(B)} - \text{type inhibitory local control} \\ & + \text{(C)} - \text{type top} - \text{down control from R2 of next} \\ & \quad \text{higher subsystem} \\ & = 1.0 \\ & + \sum_j [R1_{(n)}[k][y][x][j](t-1)]^2 - \|I_{(n-1)}[y][x](t-1) \\ & - \text{Mean}_{(n)}[k]^2 \\ & - \max_{kk \neq k} R2_{(n)}[kk][y][x](t-1) + \|P_{(n+1)}[k][y][x](t-1)\| \end{aligned}$$

3.3. Face recognition in the hierarchic modular artificial neural system

The recognition of face images is simulated as an example of hypothetical reasoning in artificial neural systems. We first describe how the system works, and then show details of the results.

3.3.1. *Time course of hypothesis generation and verification in the system.* We illustrate the overall information flow in our simulation system performing the recognition task. The related figures (Fig. 8(a)–(d)) follow a time sequence. In the simulation system, V4 subsystem is designed as a simple feedforward filtering so that the representation in V4 is not separated. However, the representations in PIT and AIT are separated into "input description" and "hypothesis description": pictures in the lower parts of PIT and AIT are "input description" only at those cortical columns where $R0$ (hypothesis generator) submodules are activated. Pictures in the upper parts of PIT and AIT are "hypothesis descriptions" where small thick squares represent the established hypotheses and small thick squares with a cross represent the erroneous hypotheses.

At time-step 1 [see Fig. 8(a)], the scene is input and the foveated region is passed to V4. V4 performs general PCA (see Appendix A) filtering with overlapping receptive fields and produces first level input description. V4 input description is fed forward to $R0$ of AIT via a shortcut pathway to define the initial global hypothesis ("human face") out of many alternative feature classes. This hypothesis is in turn fed back to V4 to segment a processing area as shown in Fig. 8(a). V4 input description is also fed forward to $R0$ of PIT which performs template matching. Those $R0$ of PIT that signal the presence of oriented contours are activated first because of the higher matching scores compared with $R0$ of PIT that signal the presence of other local features. At those PIT columns where $R0$ are activated, the hypotheses are described in terms of the respective feature-specific eigenspace in $R1$ and are verified by DFFS in $R2$. In Fig. 8(a), all the hypotheses are successfully verified and feedforward hypotheses are established in PIT. The established hypotheses of oriented contours are fed forward to AIT to add more details to the input description.

At time-step 2 [see Fig. 8(b)], those inputs from V4 to PIT which previously activated $R0$ of PIT with oriented

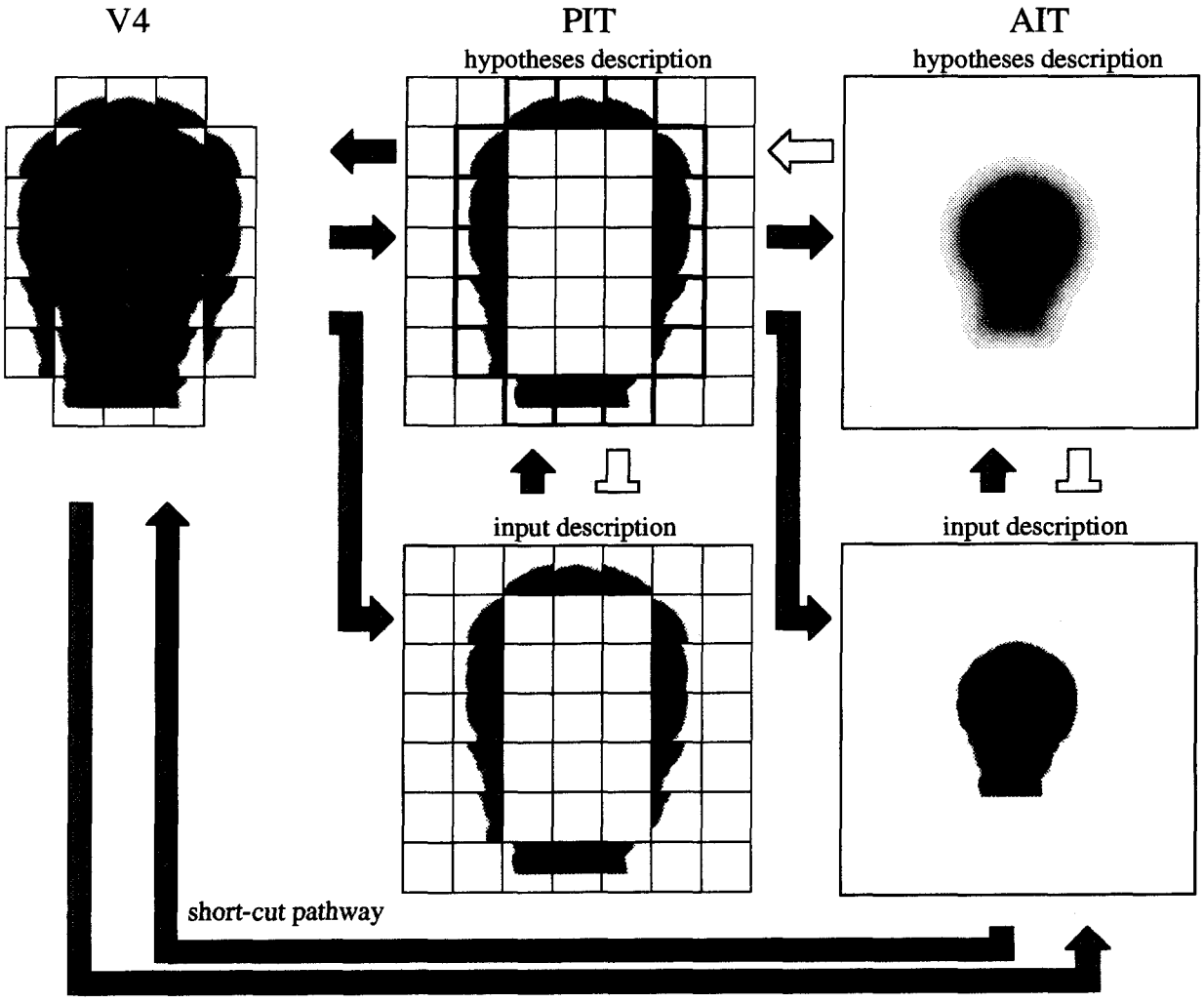

FIGURE 8. Time course of hypothesis generation and verification.

contours are switched-off by the inhibitory feedback from *R*2 of PIT, so that *R*0 of PIT with matching inner facial parts (eye, nose, mouth) get the chance to be activated. Since the recognition ability of *R*0 template matching is relatively low, there exist some erroneous hypotheses. Again, at those PIT columns with activated *R*0, the hypotheses are described in terms of the respective feature-specific eigenspace in *R*1 and are verified by DFFS in *R*2. By this verification, three correct hypotheses are established (right eye, left eye, nose) and erroneous hypotheses are rejected (features at the forehead). However, there is still one missing hypothesis (mouth) in PIT, since the ability of pure local processing is limited and the global feedback from *R*1 and *R*2 of AIT is not yet definite at this time. Note that different from input description, previously (at time-step 1) established hypotheses of oriented contours are still kept in hypothesis description of PIT. The established hypotheses of inner facial parts are fed forward to AIT and a somewhat elaborate hypothesis ("specific type of human face") emerges in *R*1 of AIT.

At time-step 3 [see Fig. 8(c)], with the definite global feedback from *R*2 of AIT, the erroneous hypothesis at time-step 2 is rejected and the missing hypothesis (mouth) is generated and verified in PIT. With these established local features, the AIT global hypothesis becomes more and more specific.

The interpretation process stops when no more strong activation remains in *R*0 of PIT [see Fig. 8(d)]. At this final state, all the facial parts are represented correctly in PIT.

3.3.2. *Results.* We first demonstrate the local hypothesis generation and verification at PIT in Fig. 9. In Fig. 9(a), where the predictive feedback from AIT to PIT is cut, the system cannot generate a reliable local hypothesis for the mouth. In Fig. 9(b), where the predictive feedback is performed, the system can generate a reliable local hypothesis for the mouth at time-step 4. These processes are analyzed in detail in Fig. 10, in which the activation of the local hypotheses of facial parts at a position where a "mouth" is located in the sensory input and the

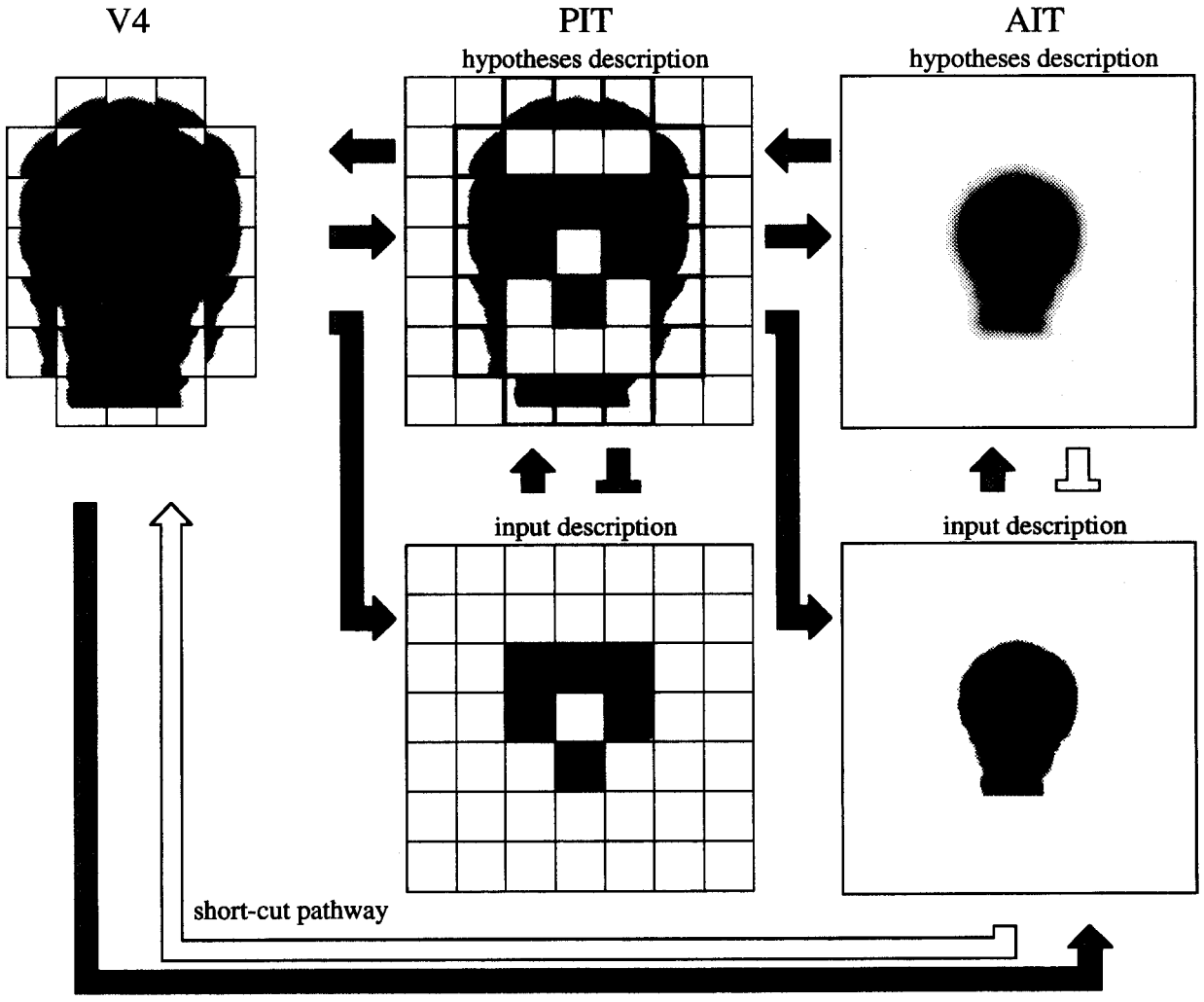

FIGURE 8. Continued.

distance from a specific face are shown. Figure 10(a) refers to Fig. 9(a), and Fig. 10(b) to Fig. 9(b). With Fig. 10(a), the local hypothesis of "mouth" is completely silent. On the other hand, an incorrect hypothesis of "left eye" becomes active at time-step 3 and suppresses other hypotheses. With Fig. 10(b), the local hypothesis of "mouth" is activated at time-step 3 by the predictive feedback, and $R2$ of "mouth" accepts this hypothesis at time-step 4. The incorrect hypothesis of "left eye" is suppressed by the predictive feedback at time-step 3. The system with feedback has a superior performance for local feature detection (see Table 1).

Secondly, we demonstrate the effect of this feedback on the identification of faces. The lowest row in Fig. 10 shows the distance from a specific face in feature space (identification of face). At time-step 1, both systems, with and without predictive feedback, roughly recognize a correct face without detecting all of the facial features at PIT. At time-step 2, the systems roughly find the position of a face by detecting the positions of some local facial features at PIT and move the processing area, so that both systems increase the discriminability. From time-step 3, only the system of (b) can increase discriminability more by detecting a local feature "mouth".

3.3.2.1. Effect of predictive feedback. The state of predictive feedback at time-step 2 in Fig. 9(b) shows that the AIT rapidly and coarsely recognizes the object as a face and applies the correct hypothesis of the configuration of local features at PIT. That hypothesis not only lets the system find even a deformed local feature at the expected position, but also accelerates the recognition. Even if some facial features are removed, AIT will generate the hypothesis of a face from the global configuration, but may not generate the hypothesis of a known specific person. By refining the hypothesis, AIT detects the correct position of local facial features at PIT, and obtains the correct local information from PIT. Finding the correct position of local facial features increases the identification rate.

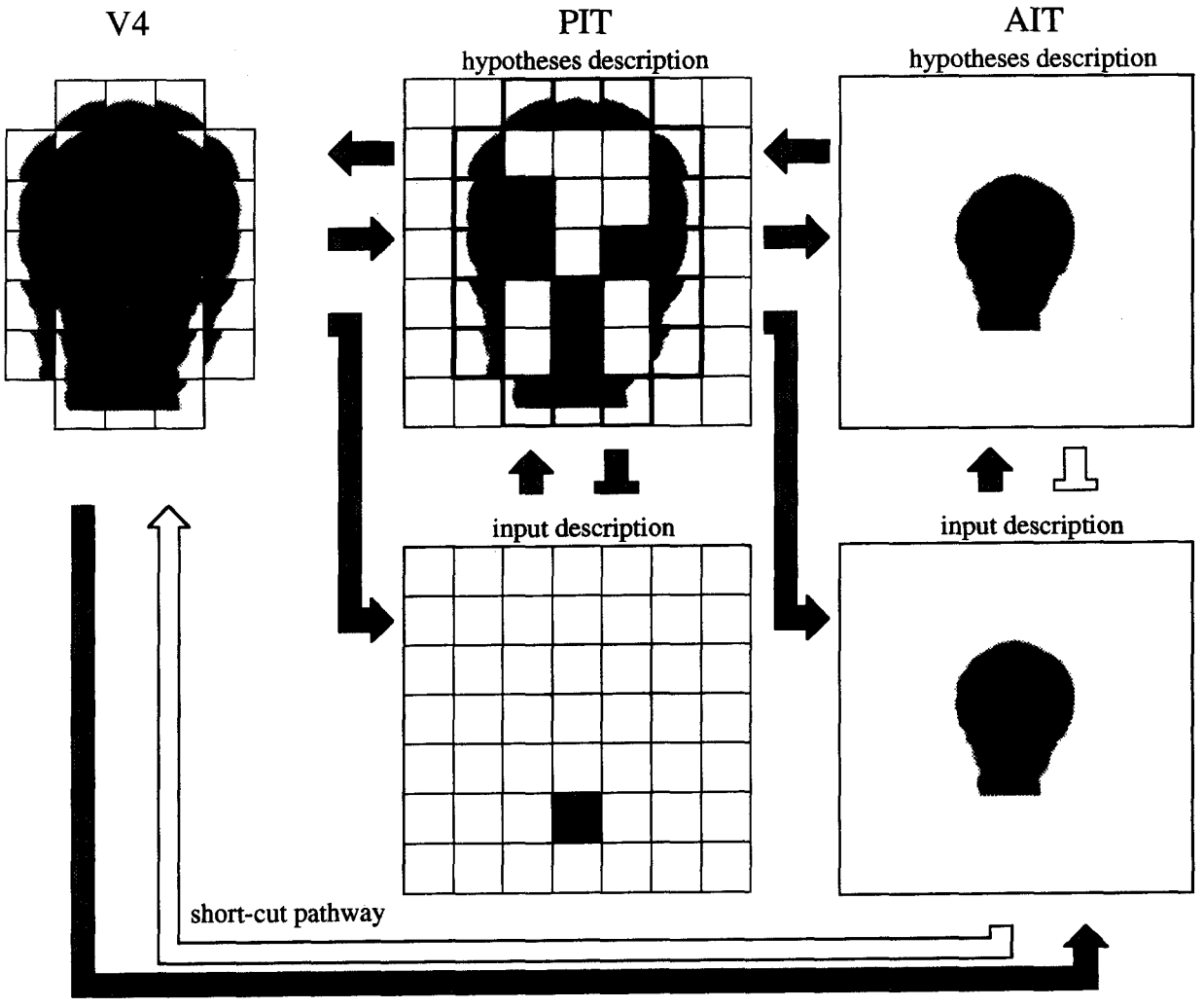

FIGURE 8. Continued.

3.3.2.2. Effect of separation of input description and hypothesis

Separation of input description and hypothesis is crucial for identification of a certain person's face in the process to match the input description with the best fitting recalled knowledge. In any case, having a small number of relatively simple objects described by salient local features without variations and fluctuations, several

TABLE 1
Local feature detection rates

Rates	A with feedback	B without feedback	C Eigenfeature	D MLP
Hit rate (eyes)	99.5% (209/210)	94.3% (198/210)	94%	93.2% (41/44)
FA rate (eyes)	0.5% (1/210)	5.7% (12/210)	6%	
Hit rate (nose)	100.0% (105/105)	92.4% (97/105)		
FA rate (nose)	0.0% (0/105)	7.6% (8/105)		
Hit rate (mouth)	99.0% (104/105)	76.2% (80/105)		
FA rate (mouth)	1.0% (1/105)	23.8% (25/105)		

Comparison between the proposed method and others. Method A is the proposed method in this paper. We show in comparison the results of method B performed with conditions identical to method A, but missing predictive feedback. Method C is the eigenfeature method. Method D is based on the eigenfeature and DFFS using a multilayer perceptron (MLP). The feature detection ability is measured for four facial features by method A and B where face images are sampled from 35 persons with three different shots. Each face in the sample is upright, similar size and the background is removed. Thirty-five faces from a set of shots are used for training AIT of the system. The detection rates of method C and D were verified for the eye of different sets of samples, and citing Pentland et al. (1994) and Waite (1992), respectively. Hit rate is the probability that we detect the feature when it is present. False alarm (FA) rate is the probability that we detect the feature when it is not present or we do not detect anything when it is present.

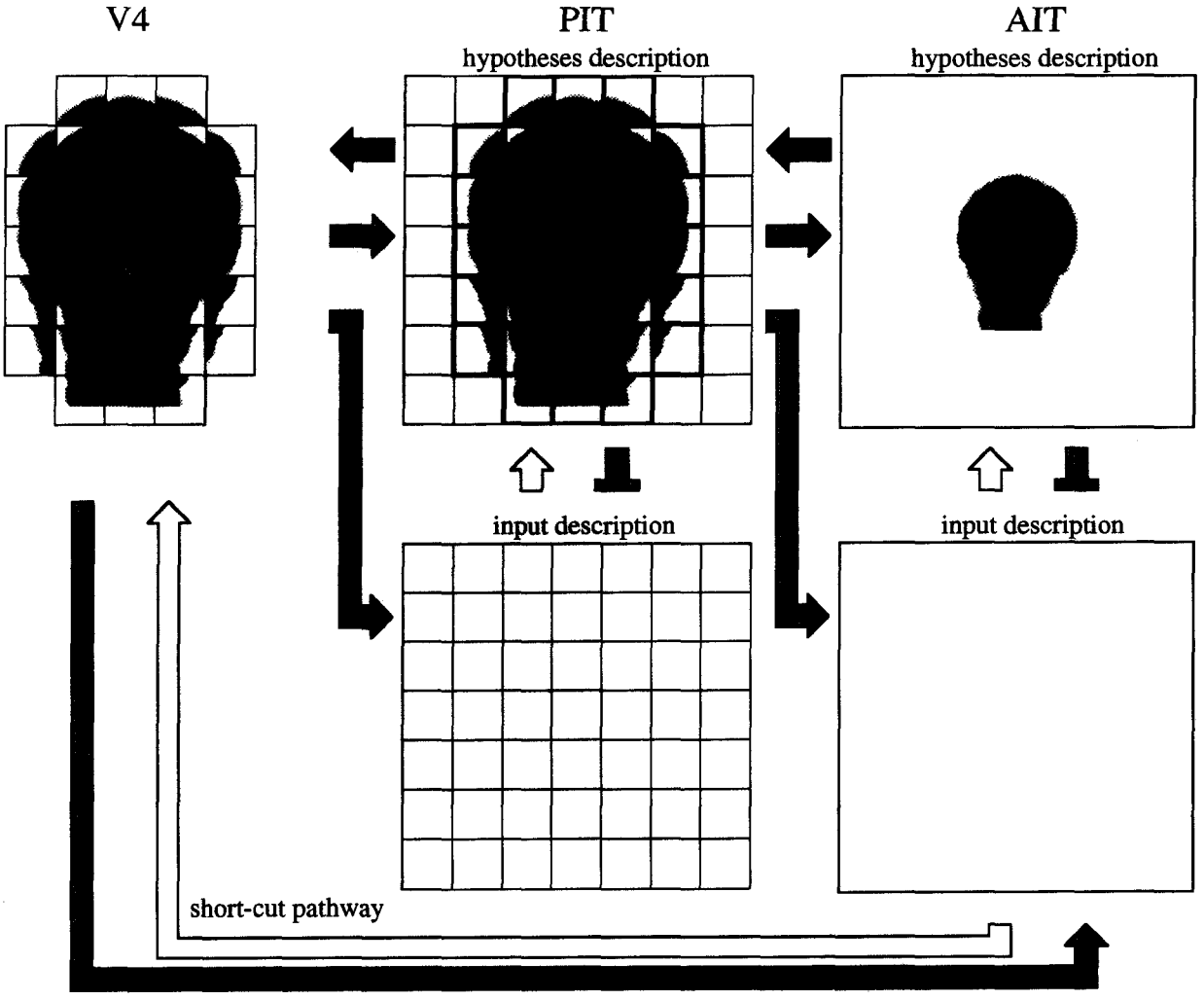

FIGURE 8. Continued.

alternative hypotheses on the input will appear. Modulating the input description would deprive the system of the possibility to evaluate the distance of the emerging hypothesis to the sensory reality and to update acquired knowledge or change the hypothesis (see Section 1.2 and Section 2.3). The course of local feature activation [Fig. 9(b)] shows that this separate representation enables the parallel evaluation of the hypotheses and the adaptive limitation of the breadth of search. The time course of the distance measure [Fig. 10(b)] shows that the coarse hypothesis description at $R1$ is continuously converging to the refined hypothesis description. The hypothesis description must be of a compact quality to have the capability of both representing variations of the features not seen before and enabling such a smooth convergence of the decision process. Despite being a controversial issue with regard to modeling neural encoding, PCA can simulate that required flexible description.

4. DISCUSSION

The neocortex should be looked at as the system that makes use of the acquired knowledge to decompose the stream of sensory data into (for the system) meaningful pieces, and to eliminate the expected pieces from the input description, isolating those aspects of the input not yet accounted for by the generated internal description (Barlow, 1994; Koerner, 1994). That the cortex should perform such a type of processing has been stressed recently (Barlow, 1994; Churchland et al., 1994; Koerner, 1994; Mumford, 1994). In this paper we propose the architecture that can implement this quality of processing. We tried to define a general architecture by "abducting" crucial characteristics of the neocortical type of processing and of its neural hardware organization which, as we claim, is different to lower-level types of signal filtering and approximation implemented well by formal neural network architectures composed of simple formal neurons as elementary processing nodes. Different description levels for encoding of information flow and its inherent control must be reflected by a respective proper neural network architecture too. Signal filtering and approximation in lower sensory processing stages are surely different in algorithmic

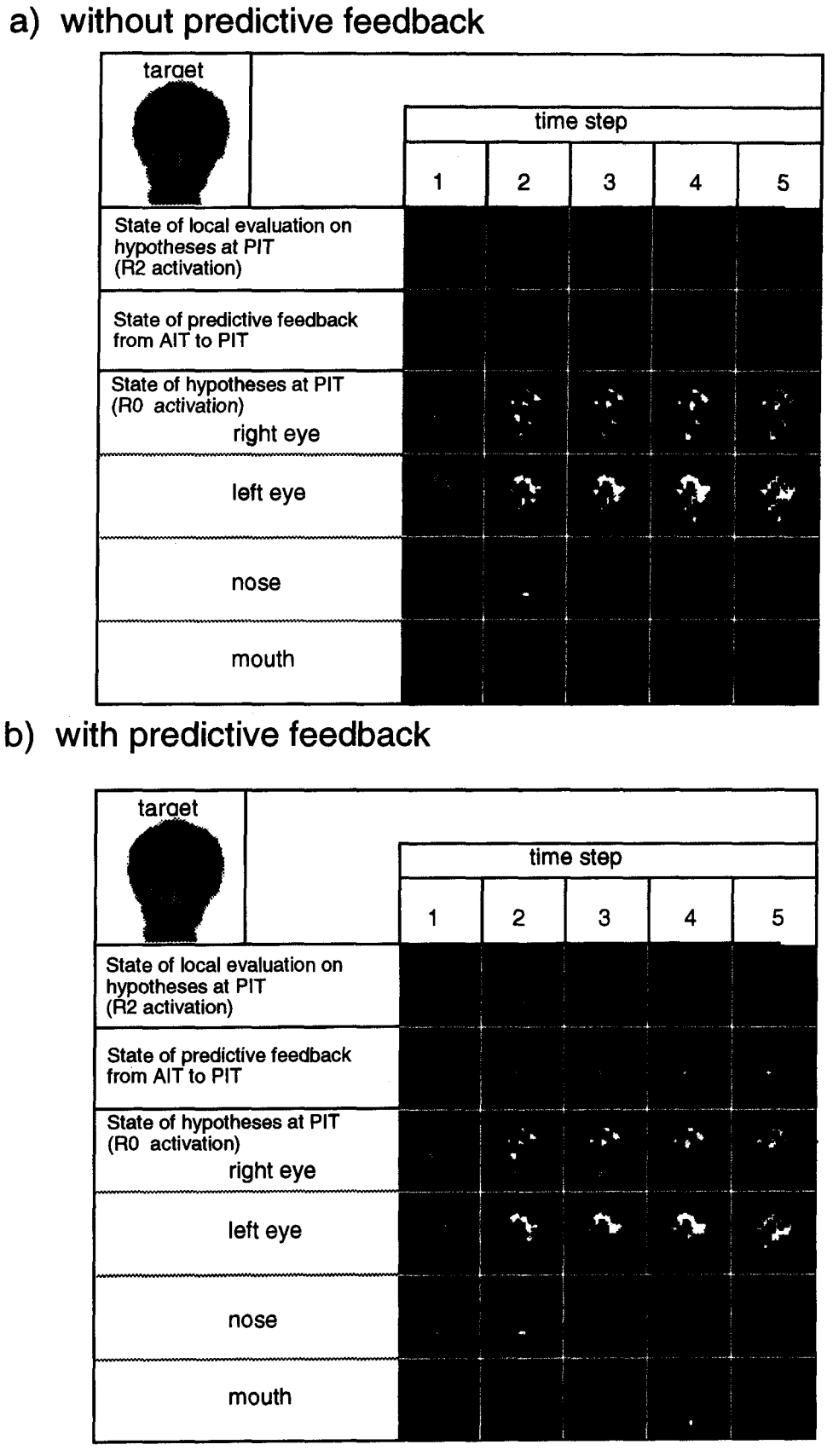

FIGURE 9. The local hypothesis generation and verification at PIT. The target image is shown at the upper left corner and time runs from left to right. The first row shows the state of predictive feedback from AIT to PIT, next row shows the state of local verification (*R2* activation), and the next four rows show the state of the local hypotheses (*R0* activation) of four feature classes: right eye, left eye, nose and mouth. The position of each activation is the same as in the target image.

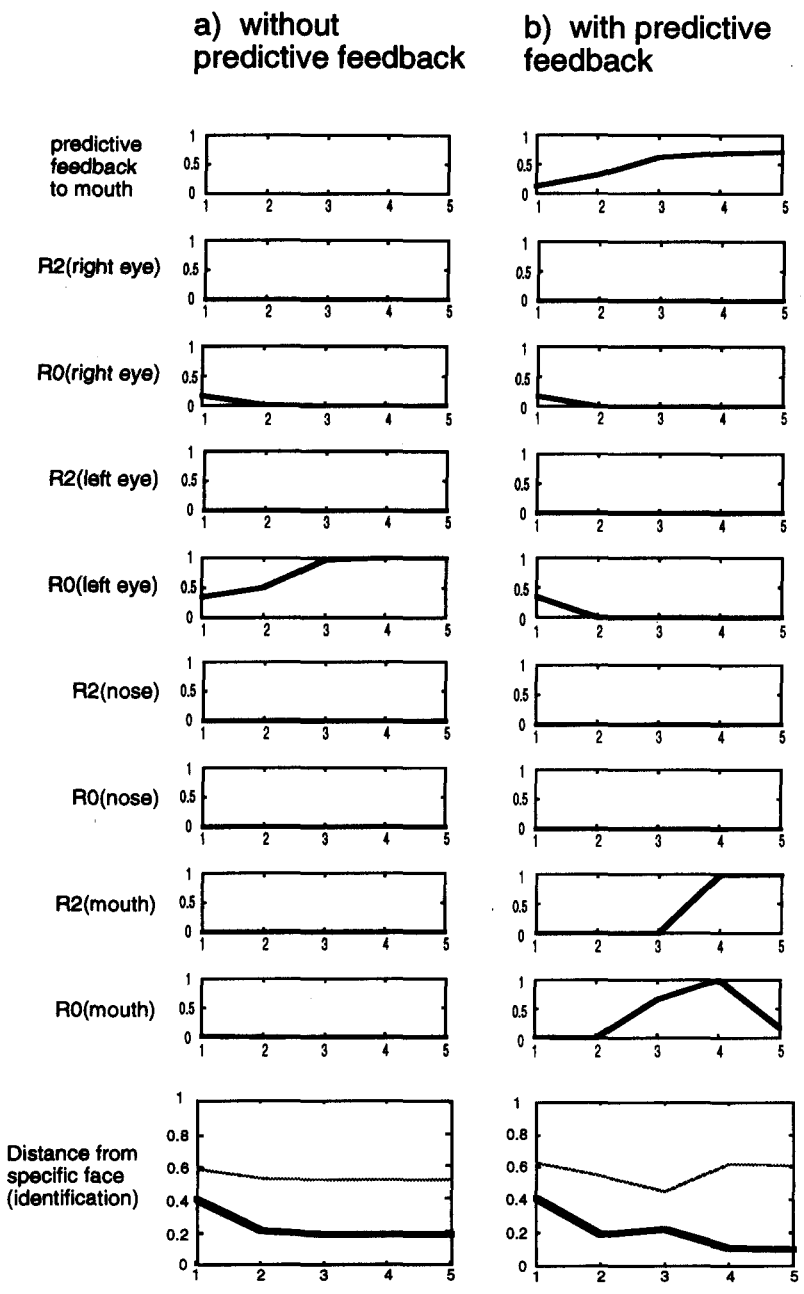

FIGURE 10. The activation of the local hypotheses. Time runs from left to right, and is the same as Fig. 9. The first row shows the active feedback from AIT to PIT for "mouth" at this location. The next rows show *R2* and *R0* activation of the respective feature classes: *right eye, left eye, nose and mouth. The lowest row shows the distance from a specific face in feature space* (identification of face), where the thick line indicates the distance from a correct face and the thin line indicates that from the second nearest face. The number of memorized faces in the system is 35 which is used for developing AIT representation.

structures compared with internal simulation at different abstraction levels, as we propose for the cortical processing stages. We limit our considerations here to a simple visual problem, but the crucial feature is the modular architecture that implements this type of reasoning. Rich recurrence in visual processing means that stored information from earlier learning plays a role in what the living being "sees". In the ventral visual pathway there is not only a point-to-point recurrence between subsequent stages of processing, but the feedback even contacts from IT directly to V4, V2, V1 and even LGN (Rockland et al., 1992). Setting a predictive feedback from IT to the sensory input of the visual cortex does not need several iterations down the hierarchy, but may be achieved using only a very few. If the input is well known and of good resolution, the recognition resembles almost a parallel recall. For more fuzzy or less known aspects in the input, a longer sequence of iterative refinement of the

decision by reasoning will occur. We want to stress here that the system does not provide a decision beyond the verifiable refinement of its hypothesis. If no further refinement is possible, the process comes to a halt at the level of resolution that can be reliably verified. Note, that the current simulation system is only a limited implementation of our model framework for robust recognition and autonomous learning as proposed in Section 2. Learning is not yet implemented, and together with the self-organization of invariant knowledge representation in the artificial neural system, the problem of knowledge-based top-down segmentation has to be solved to work on real world input without constraining conditions.

Since we consider image recognition to be one of the problems that should be implemented in this neocortical type of hardware for hypothetical reasoning, our theory and simulation should enable predictions on the course of recognition in the ventral visual pathway. We focus here on the following problems that are currently under discussion in experimental brain science: visual recognition as a forward or an iterative process; visual segmentation by synchronization; and the smooth transition from signal to symbol processing.

4.1. Visual object recognition is not a pure forward process

Face selective cells at IT cortex respond with a latency of about 100 ms, which is barely several milliseconds more than the time required for signal transmission from retina to IT cortex. Since the sustainable maximum firing rate of cortical neurons is about 100 Hz only, it has been concluded that recognition in the ventral visual pathway is a pure feedforward process, probably mediated by the very first spike transmitted by any neuron in the hierarchy upward to IT cortex — hence, there is no time for iteration (Thorpe & Imbert, 1989; Oram & Perrett, 1994; Thorpe et al., 1996). So are the rest of the spikes in the 200 ms duration of striate neurons nothing but noise? Not necessarily — we contest that assumption. We do not deny that forward processes are instrumental in triggering the recognition process, but we claim that this is only one part of the full recognition story which involves both the forward activation of a coarse hypothesis and its iterative refinement by the proposed hypothetical reasoning. In vision, three-dimensional objects in the environment are compressed into two-dimensonal images on the retina, a many-to-one mapping with loss of information. Objects in the visual environment are highly variable and non-stationary. Recognition of real world objects from this visual input cannot be properly solved without constraints given beforehand (Marr, 1982; Poggio et al., 1985; Kawato et al., 1993). We propose that without an established context that feeds back the proper constraints by activating an expected representation, recognition must be a multistage process: coarse activation of a behavioral evaluation of the visual input via the shortcut through amygdala; selecting one class of objects out of the set of primed alternatives by express forward descriptions of salient features, which is a very sparse subset of all activated forward descriptions and can be sufficiently distinctive to activate this one hypothesis only; feedback of the hypothesis to set the object-shaped segmented area properly controlling the next arriving less salient object descriptions while suppressing the barrage of unrelated input of the background that would only de-stabilize the recognition process. The number of iterations should be limited in most cases: the time between eye saccades is in the range between 200 and 300 ms, and this time window nicely correlates with the measured response duration of striate cortex neurons (Dinse, 1994; Freeman, 1995). In the case of continuous recognition, the previously fully developed internal image (detailed hypothesis) of the scene may be compared against a slightly changing input, enabling the evaluation of even complex scenes in the 100 ms range — without the need for many iterations. In support of our prediction is the observation that neurons in superior temporal sulcus with low selectivity have short response duration, while neurons with high selectivity have longer response duration and a longer latency in response onset (Mikami et al., 1994). Surround effects in V1 neurons are not observed during the first 8 ms of the response, but are built up progressively thereafter if the stimulus also covered the surrounding receptive field of the neuron (Knierim & Van Essen, 1992).

The experiments that gave rise to the "only forward processing counts" assumption may actually not support that assumption sufficiently. The definite activation of a face coding neuron in IT cortex may be explained by the priming of a limited set of hypotheses via amygdala, and subsequent selection of one out of this set by the limited set of express salient features arriving before the bulk of afferent activity. Moreover, the experimental conditions could have facilitated expectations that limited the decision problem considerably, as a limited number of objects were presented repeatedly to the monkey.

Human subjects can determine whether or not an animal is within a randomly selected natural scene flashed for only 20 ms on the screen (Thorpe et al., 1996). In this experiment, the task description to look for an animal already gives some constraint, which may be focused by the amygdala shortcut to the extent that some salient features could trigger the internal image "animal x". To verify our theory it is important to assess what the subjects indeed recognized, and what they subsequently generated by imagery. For instance, asking human subjects confronted with a complex scene for an interval between 30 and 200 ms (to exclude eye movements) to report what they have recognized and to describe as much detail as possible, may reveal the refinement process of the recognition with time. A better controlled experimental condition of a go/no-go

task for this purpose is the following setup: subjects are presented with an object occluded to some part by a white plane except at those positions where the most salient features of this object are located (like a Swiss cheese). The holistic activation by the shortcut via amygdala may be ineffective under this condition, but the limited number of salient features should enable the forward express signaling to activate the correct representation at IT cortex. Adding next distractors of the same saliency randomly over this Swiss-cheese like cover plane should create a markedly longer latency for recognition.

4.2. Tuning-in of the module's response

The activation of coarse hypotheses at *R*0, and its refinement by *R*1 with subsequent WTA decision making is a dynamic tuning-in of the feature filters mediated by excitatory cortical feedback from the upper layers of the same column, as observed by Tommerdahl et al. (1993). The refinement-based tuning-in supported by feedback is a plausible model for the modulation of the tuning characteristics of feature filters in the first visual area as reported by Shevelev et al. (1993) and Dinse (1994). According to our theory, this is not a contradiction to the report that feature filters are tuned sharply from the onset of the afferent input and do not change their response characteristics (Celebrini et al., 1993; Oram & Perrett, 1992; Rolls & Tovee, 1994). Our theory predicts that feature filters that have salient features in their receptive field will be forward activated with all the available resolution, while those without may need predictive feedback from the emerging higher level hypothesis to resolve the local decision problem. Instead of testing the tuning with salient stimuli, line-based figures that have locally fuzzy distortions or missing parts should be used in experiments to prove this prediction. The "illusory contour detectors" recently reported in V1 and V2 (von der Heydt et al., 1984; Grosof et al., 1993; Davis & Driver, 1994) could well be a phenomenon related to the predictive feedback by a hypothesis such as we propose in our theory. The measurement of the relative timing of the response onset (the very first spikes) of each feature detector neuron should verify that point. Our prediction is supported by observations that illusory contours disappear completely after fixation for 4 s (long enough to enable the refinement of the hypothesis in detail), and persist for a surprising long time (300 ms) after the inducing elements have been switched off (Ramachandran et al., 1994).

4.3. Synchronization of cortical oscillatory activity and top-down control

A widely accepted hypothesis (Singer, 1993) proposes that active synchronization between populations of visually driven neurons could be the basis for visual segmentation and perceptual grouping. If coherent afferent activation would indeed define this segmentation, then temporal asynchrony of parts of a visual object in presentation to human subjects should effect significantly the perception. However, it has been demonstrated that performance of the subjects in recognition of the features or objects was neither helped nor hindered by the imposed temporal synchronicity, and the conclusion was that not synchronicity but small local displacement appears to be essential for binding (Keele et al., 1988; Fahle & Koch, 1995; Kiper et al., 1996).

Our theory provides a reasonable explanation and prediction for this phenomenon. Hypotheses of features/objects emerge at that level of representation where the receptive field of its modular units covers the full spatial arrangement of the stimulus at the retina. Local features may be bound together into synchronous ensembles of V1 or V2 neurons by the emerging higher level hypothesis and its feedback, but not by the proposed active feature linking. For a top-down defined object at the retina, small temporal asynchrony does not matter, but relative spatial displacement of parts of the object does matter.

4.4. Smooth transition from signal to symbol processing

Highly abstracted representations as they are to be expected at the cortical level must consist of at least two parts: an invariable part which defines the class of the representation, and a variable part (default options) which describes details and variations in the represented object. We allocated the invariable part to the activation of a columnar unit (a module in our artificial neural system architecture), and the variable part to the activation of a distinctive set of neurons of that module (Koerner et al., submitted). We consider the single module to represent the neural equivalent of a symbol: the same symbol can represent a class (in the case of coarse activation), or a singular highly specific object out of that class (very distinct activation pattern within the module in the case of definite input description and complex intracortical activation). Both types of symbolic description must potentially exist at the same time, and the symbol may be instantiated in any specific configuration tolerable by the class definition encoded in the module. Because of the compact representation within the module, the symbol can be instantiated to describe the sensory input even in a configuration never explicitly experienced by the system before. Hence, the proposed modular architecture seems to be a proper implementation of a "neural symbol description" which incorporates both the classical logical and signal description.

REFERENCES

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following damage to the human amygdala. *Nature*, **372**, 669–672.

Allison, J. D., Casagrande, V. A., & Bonds, A. B. (1995). The influence of input from the lower cortical layers on the orientation tuning of upper layer V1 cells in a primate. *Visual Neuroscience*, **12,** 309–320.

Ando, H. (1996). 3D Object Recognition Using Bidirectional Modular Networks. In S. Z. Li, D. P. Mital, E. K. Teoh, H. Wang (Eds.), *Recent Developments in Computer Vision*, (ACCV'95) Invited Session Papers) (pp. 467–478). Berlin, Heidelberg, New York: Springer-Verlag.

Barlow, H. (1994). What is the computational goal of the neocortex?. In C. Koch & J.L. Davis (Eds.), *Large-scale theories of the brain* (pp. 1–22). Cambridge, MA: MIT Press.

Barnden, J.A. (1995). Artificial intelligence and neural networks. In M.A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 98–102). Cambridge, MA and London, England: MIT Press.

Brunelli, R., & Poggio, T. (1993). Face recognition: features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15,** 1042–1052.

Cauller, L. J., & Connors, B. W. (1994). Synaptic physiology of horizontal afferents to layer 1 in slices of rat SI cortex. *Journal of Neuroscience*, **14,** 751–762.

Cavanagh, P. (1991). What's up in top-down processing? In A. Gorea (Ed.), *Representations of vision* (pp. 295–304). Cambridge, UK: Cambridge.

Celebrini, S., Thorpe, S., Trotter, Y., & Imbert, M. (1993). Dynamics of orientation coding in area V1 of the awake primate. *Visual Neuroscience*, **10,** 811–825.

Churchland, P., Ramachandran, V.S. & Sejnowski, T.J. (1994). A critique of pure vision. In C. Koch & J.L. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 23–60). Cambridge, MA: MIT Press.

Davis, G., & Driver, J. (1994). Parallel detection of Kansiza subjective figures in the human visual system. *Nature*, **371,** 791–793.

Dinse, H.R. (1994). A time-based approach towards cortical functions: neural mechanisms underlying dynamic aspects of information processing before and after postontogenetic plastic processes. *Physica D*, **75,** 129–150.

Fahle, M., & Koch, C. (1995). Spatial displacement, but not temporal synchrony, destroys figural binding. *Vision Research*, **35,** 491–494.

Felleman, D.J., & VanEssen, D.C. (1991). Distributed hierarchical processing in primate visual cortex. *Cerebral Cortex*, **1,** 1–47.

Freeman, R.D. (1995). Space and time in the central visual pathways. *Proceedings of 5th International Symposium on Bioelectronic and Molecular Electronic Devices*, Okinawa, Japan, B1-01.

Fujita, I., & Fujita, T. (1996). Intrinsic connections in the macaque inferior temporal cortex. *The Journal of Comparative Neurology*, **368,** 467–486.

Fukushima, K. (1986). A neural network model for selective attention in visual pattern recognition. *Biological Cybernetics*, **55,** 5–15.

Gregory, R.L. (1995). Seeing backwards in time. *Nature*, **373,** 21–22.

Grosof, D. H., Shapley, R. M., & Hawken, M. J. (1993). Macaque V1 neurons can signal "illusory" contours. *Nature*, **365,** 550–552.

Hecht-Nielssen, R. (1995). Replicator neural networks for universal optimal source coding. *Science*, **269,** 1860–1864.

Hernandez-Gonzales, A., Cavada, C., & Reinoso-Suarez, F. (1994). The lateral geniculate nucleus projects to the inferior temporal cortex in the macaque monkey. *Neuro Report*, **5,** 2693–2696.

Hinton, G. (Ed.) (1991). *Connectionist symbol processing*. Cambridge, MA: MIT Press.

Jeannerod, M. (1995). Mental imagery in the motor context. *Neuropsychologia*, **33,** 1419–1432.

Kawato, M., Hayakawa, H., & Inui, T. (1993). A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network*, **4,** 415–422.

Keele, S. W., Cohen, A., Ivry, R., Liotti, M., & Yee, P. (1988). Tests of temporal theory of attentional binding. *Journal of Experimental Psychology*, **14,** 444–452.

Kiper, D.C., Gegenfurther, K.R., & Movshon, J.A. (1996). Cortical oscillatory responses do not affect visual segmentation. *Vision Research*, **36,** 539–544.

Knierim, J. J., & Van Essen, D. C. (1992). Neuronal responses to static texture patterns in area V1 of the alert monkey. *Journal of Neurophysiology*, **67,** 961–980.

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, **71,** 856–867.

Koerner, E. (1994). Autonomous recognition and selforganization of knowledge representation in neural networks: Part 1: From structuring data to constraint generation by self-referential control. *Holonics*, **4,** 3–34.

Koerner, E., Tsujino, H. & Masutani, T. (submitted). A model of symbol/signal representation in the primate ventral visual pathway. *Neural Networks*.

Kosslyn, S. M., Alpert, N. M., Thompson, W. L., Maljkovic, V., Wise, S. B., Chabris, C. F., Hamilton, S. E., Rauch, S. L., & Buonanno, F. S. (1993). Visual mental imagery activates topographically organized visual cortex: PET investigations. *Journal of Cognitive Neuroscience*, **5 (3),** 263–287.

Kosslyn, S. M., Thompson, W. L., Kim, I. J., & Alpert, N. M. (1995). Topographical representation of mental images in primary visual cortex. *Nature*, **378,** 496–498.

Leopold, D. A., & Logothetis, N. K. (1996). Activity changes in early visual cortex reflect monkey's percepts during binocular rivalry. *Nature*, **379,** 549–553.

Marr, D. (1982). *Vision*. San Francisco: Freeman Publishers.

Mikami, A., Nakamura, K., & Kubota, K. (1994). Neuronal responses to photographs in the superior temporal sulcus of the rhesus monkey. *Behavioral Brain Research*, **60,** 1–13.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), *The psychology of computer vision* (pp. 211–277). New York: McGraw-Hill.

Minsky, M. (1985). *The society of mind*. New York: Simon & Schuster.

Mumford, D. (1992). On the computational architecture of the neocortex II. The role of cortico-cortical loop. *Biological Cybernetics*, **66,** 241–251.

Mumford, D. (1994). Neuronal architectures for pattern-theoretic problems. In C. Koch & J.L. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 125–152). Cambridge, MA: MIT Press.

Oram, M. W., & Perrett, D. I. (1992). Time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology*, **68,** 70–84.

Oram, M. W., & Perrett, D. I. (1994). Modeling visual recognition from neurobiological constraints. *Neural Networks*, **7,** 945–972.

Pentland, A. P., Moghaddam, B., & Starner, T. (1994). View-based and modular eigenspace for face recognition. *Proceedings of Computer Vision and Pattern Recognition '94* (pp. 84–91).

Picton, T. W., & Stuss, D. T. (1994). Neurobiology of conscious experience. *Current Opinion Neurobiology*, **4,** 256–265.

Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature*, **317,** 314–319.

Poggio, T. & Hurlbert, A. (1994). Observations on cortical mechanisms for object recognition and learning. In C. Koch & J.L. Davis (Eds.), *Large-scale theories of the brain* (pp. 151–182). Cambridge, MA: MIT Press.

Ramachandran, V. S., & Cobb, S. (1995). Visual attention modulates metacontrast masking. *Nature*, **373,** 66–68.

Ramachandran, V. S., Ruskin, D., Cobb, S., Rogers-Ramachandran, D., & Tyler, C. W. (1994). On the perception of illusory contours. *Vision Research*, **34,** 3145–3152.

Rao, R. O. N., & Ballard, D. H. (1995). *Dynamic model of visual memory predicts neural response properties in the visual cortex*. Technical Report 95.4, Department of Computer Science, University of Rochester.

Rockland, K. S., Saleem, K., & Tanaka, K. (1992). Widespread

feedback connections from areas V4 and TEO. *Visual Neuroscience*, **11,** 597–600.
Rockland, K. S., & VanHoesen, G. W. (1994). Direct temporal-occipital feedback connections to striate cortex (V1) in the macaque monkey. *Cerebral Cortex*, **4,** 300–313.
Rodman, H. R. (1994). Development of inferotemporal cortex in the monkey. *Cerebral Cortex*, **5,** 484–498.
Rolls, E. T., & Tovee, M. J. (1994). Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proceedings of the Royal Society of London*, **B 257,** 9–15.
Romanski, L. M., & LeDoux, J. E. (1992). Equipotentiality of thalamo-amygdala and thalamo-cortico-amygdala circuits in auditory fear conditioning. *Journal of Neuroscience*, **12 (11),** 4501–4509.
Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Bream, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8,** 382–439.
Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: a connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, **16,** 417–494.
Shevelev, I. A., Sharaev, G. A., Lazareva, N. A., Novikova, R. V., & Tikomirov, A. S. (1993). Dynamics of orientation tuning in the cat striate cortex neurons. *Neuroscience*, **56 (4),** 865–876.
Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology*, **55,** 349–374.
Smith, J. D., Wilson, M., & Reisberg, D. (1995). The role of subvocalization in auditory imagery. *Neuropsychologia*, **33,** 1433–1454.
Smolenski, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, **11,** 86–116.
Tanaka, K., Saito, H., Fukuda, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, **66 (9),** 170–189.
Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science*, **262,** 685–688.
Thorpe, S. J., & Imbert, M. (1989). Biological constraints on connectionist models. In E. Pfeifer, R. Schreter, Z. Fogelman-Soulie & L. Steels (Eds.), *Connectionism in perspective* (pp. 63–92). North-Holland: Elsevier.
Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, **381,** 520–522.
Tommerdahl, M., Favorov, O., Whitsel, B. L., Nahkle, B., & Gonchat, Y. A. (1993). Minicolumnar activation patterns in cat and monkey SI cortex. *Cerebral Cortex*, **3 (5),** 399–411.
Turk, M. A., & Pentland, A. P. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, **3 (1),** 71–86.
Ullman, S. (1995). Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex*, **5,** 1–11.
von der Heydt, R., Peterhans, E., & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science*, **224,** 1260–1262.
Waite, J. B. (1992). Training multilayer perceptrons for facial feature location: a case study. In R. Linggard, D. J. Myers & C. Nightingale (Eds.), *Neural networks for vision, speech and natural language* (pp. 30–49). London, UK: Chapman & Hall.
Wiskott, L., & von der Malsburg, C. (1995). Recognizing faces by dynamic link matching. *Proceedings of International Conference on Artificial Neural Networks '95*, Paris, France, pp. 347–352.

APPENDIX A: CONFIGURATION OF THE SUBSYSTEM

The detailed configuration of the subsystem of the simulation is described here. The notations refer to the equations in Appendix 3.2.2.

- *V4-subsystem: local feature I.* We define one class ($K = 1$) for the modules in V4, so we omit $R0$ at V4. The only class mainly represents the spatial frequency of gray-level local input image. The input description of a module in V4 is 9×9 pixels local image. To prepare the input description for modules in PIT which has about a three times larger receptive field than those in V4, we simplified V4 by replacing the size of the input description from 9×9 to 27×27. $PC[k][j]$ of $R1$ is defined as "the general PCA" that is configurated by the eigenvectors calculated by PCA on a large number of natural gray-level images (27×27), and has a rich description for any pattern by a small number of filters. The receptive field of any V4 module on the 128×128 gray input image ($Y = 128$, $X = 128$) is locally described by analog activation of 200 eigenvectors ($J = 200$) of "the general PCA". The local verification of $R2$ at V4 is not important because of the generality of "the general PCA", but the predictive feedback from PIT and AIT to $R2$ in V4 is important, and modifies the shape of the attentional window to an increasingly fine segmentation of the feature at the input from the background according to the activated hypothesis at the AIT subsystem.
- *PIT-subsystem: local feature II.* In PIT we set $K = 16$, four classes of facial feature (right eye, left eye, nose and mouth) and 12 orientation classes. $J = 20$, $Y = 42$, $X = 42$. The input description of a module in PIT is a 200-dimensional analog activation vector from one of 3×3 respective modules in V4. *template*$[k]$ of $R0$ is defined as the mean input vector of each specific feature. $PC[k][j]$ of $R1$ is defined as the eigenvectors calculated by PCA on 105 local images sampled from three different shots of 35 persons.
- *AIT-subsystem: global feature.* We set $X = Y = 1$, since the modules of AIT cover all the attentional windows as the input site. We set $K = 1$ because we focus on one class only — the face. The number of components is set to $J = 20$. The input description is composed of an input from V4, $105 \times 105 \times 3$, and an input from PIT, $35 \times 35 \times 16 \times 20$. "$105 \times 105$ positions $\times$ 3 components" is considered to be the size of the segmented processing area at V4. Since the subsampling rate at PIT is three, the size of this segmented processing area at PIT is "35×35 positions $\times$ 16 feature classes $\times$ 20 components". *template*$[k]$, $k = 0$, of $R0$ is defined as the reconstructed input vector computed by the mean vector and the inverse projection of the working hypothesis description from the eigenspace to the input space. $PC[k][j]$, $k = 0$ and $0 \le j < 20$, of $R1$ are the eigenvectors calculated by PCA on 35 input vectors of 35 different faces from $R1$ at V4 and PIT.

APPENDIX B: DEFINITION OF BASIC CONTROL IN THE SIMULATION

- (A) *Excitatory local control.* The excitatory local control for $R0[k]$ at (y, x) in the nth subsystem is

$$\sum_{\text{neighbor}yy,xx} R0_{(n)}[k][yy][xx]$$

$$\equiv \sum_{y-1 \le yy \le y+1} \sum_{x-1 \le xx \le x+1} R0_{(n)}[k][yy][xx] - R0_{(n)}[k][y][x] \quad \text{where} 0 \le k < K,$$

$$0 \le x < X,\ 0 \le y < Y.$$

- (B) *Inhibitory local control.* Inhibitory local control is defined by the maximum $R0$ or $R2$ activity of the same macro-column (y,x), i.e.

$$\text{Inhibitory local control for } R0_{(n)}[k][y][x]$$

$$\equiv \max_{kk \ne k} R0_{(n)}[kk][y][x]$$

$$\text{Inhibitory local control for } R2_{(n)}[k][y][x]$$

$$\equiv \max_{kk \ne k} R2_{(n)}[kk][y][x].$$

- (C) *Top-down control.* Top-down control to a *module*$_{(n)}[k][y][x]$ is defined by the predictive feedback from $R2_{(n+1)}[l][u][v] (0 \le l < L, 0 \le u < U, 0 \le v < V)$ of the next higher subsystem which gets $R1_{(n)}[k][y][x]$ as its input. Since each module in the same

macro-column, (u,v), has the same receptive field, the receptive field for this $module_{(n+1)}[l][u][v], \forall l$, is defined as

$$I_{(n)}[u][v] : M(\geq J) \text{ dimensional vector}$$

where

$$R1_{(n)}[k][y][x][jj] \in I_{(n)}[u][v], \text{ i.e.}$$

$$I_{(n)}[u][v][m+jj] = R1_{(n)}[k][y][x][jj] \text{ for}$$

$$0 \leq \forall jj < J, 0 \leq \exists m < M - jj.$$

Then, the M-dimensional reconstructed image based on the $R1$ activity is

$$Rec_{(n+1)}[l][u][v] \equiv \sum_j PC_{(n+1)}[l][j]$$

$$\times R1_{(n+1)}[l][u][v][j] + \text{Mean}_{(n+1)}[l].$$

Finally, $P_{(n+1)}[k][y][x]$ is composed from $(m+1)$th to $(m+J)$th elements of $Rec_{(n+1)}[l][u][v]$. In our simulation, since we have only one AIT module, the predictive feedback recieved by $module_{(n)}[k][y][x]$ is simplified as

$$\|P_{(n+1)}[k][y][x]\|$$

$$\equiv \sqrt{\sum_{m+1 \leq mm \leq m+J} \left[Rec_{(n+1)}[0][0][0][mm]\right]^2}.$$

- (D) *Bottom-up control.* Bottom-up control from a $module_{(n)}[k][y][x]$ to a $module_{(n+1)}[l][u][v]$ is simplified to equal $I_{(n)}[u][v]$ as defined above.
- (E) *Switching-off control.* Switching-off control is simplified to equal $R2$ of the same module, i.e.

$$\text{switching} - \text{off control} \equiv R2_{(n)}[k][y][x].$$