Biol. Cybern. 79, 263–275 (1998)

Biological Cybernetics
© Springer-Verlag 1998

Forming independent components via temporal locking of reconstruction architectures: a functional model of the hippocampus

András Lőrincz

Department of Chemical Physics, Institute of Isotopes, Hungarian Academy of Sciences, Budapest, P.O. Box 77, Hungary H-1525, and Department of Adaptive Systems, Attila József University, Szeged, Dóm Square 9, Hungary H-6720

Received: 18 March 1997 / Accepted in revised form: 30 June 1998

Abstract. The assumption is made that the formulation of relations as independent components (IC) is a main feature of computations accomplished by the brain. Further, it is assumed that memory traces made of non-orthonormal ICs make use of feedback architectures to form internal representations. Feedback then leads to delays, and delays in cortical processing form an obstacle to this relational processing. The problem of delay compensation is formulated as a speed-field tracking task and is solved by a novel control architecture. It is shown that in addition to delay compensation the control architecture can also shape long-term memories to hold independent components if a two-phase operation mode is assumed. Features such as a trisynaptic loop and a recurrent collateral structure at the second stage of that loop emerge in a natural fashion. Based on these properties a functional model of the hippocampal loop is constructed.

1. Introduction

The hippocampal formation consists of three main subfields: the dentate gyrus and the CA3 and CA1 regions. The functional role of these areas is currently the subject of extensive research. The 'top-down approach' deals with the deterioriation of learning capabilities in hippocampal subjects. These investigations cover a broad range that includes behavioural studies after lesions to the brain of animals (Zola-Morgan and Squire 1991; Otto and Eichenbaum 1992; Zola-Morgan 1992 et al.) as well as studies related to human amnesia (Squire 1992; Squire and Knowlton 1995). The 'bottom-up approach' deals with hippocampal tissues in vitro and in vivo and targets the long-term potentiation (LTP) and long-term depression (LTD) learning features in these areas (Bliss and Lomo 1973; Abraham et al. 1986; Harris and Cotman 1986; Bliss and Collingridge 1993). There are studies in between that deal with both behavioural and in vivo electrophysiological data (see, e.g., Buzsáki 1989; Freund and Buzsáki 1996; Skaggs et al. 1996; O'Keefe and Burgess 1996 and references therein).

Over the years a variety of models of the hippocampus have been developed. A few representative examples will be mentioned below. The interested reader is referred to the literature for further information (see, e.g., Ono et al. 1996 and references therein). Models starting from detailed cell-level biophysical and physiological properties have strong predictive power, e.g., regarding synchronized bursts (Traub and Dingledine 1990). Another approach explains the phase precession properties of hippocampal cells by means of an asymmetric spreading activation model (Tsodyks et al. 1996). Based on electrophysiological data a statistical population model has also been constructed that reproduces epileptiform and non-epileptic rhythms in the CA3 slice (Érdi et al. 1997).

Computational models are fewer in number. For a review see, e.g., Gluck (1996). According to Rolls (1989, 1996), the hippocampus acts as an intermediate memory store with modifiable recurrent interconnections at the CA3 level that perform attractor dynamics. McClelland (1996) argues that it is computationally advantageous to form an intermediate store of episodic memories, and this store is the origin of the hippocampus. Carpenter and Grossberg (1993) suggest that the adaptive resonance theory (ART) models the interplay between the cortex and the hippocampus with the internal representation of the ART network describing the hippocampus, while the input stage of the network represents the neocortex.

The model detailed here is based on two prominent functional features of the hippocampus. (1) There is

Correspondence to: K.P. Krommenhoek
(e-mail: karin@mbfys. kun.nl, Fax: +31-24-354-1435)

Present address:
Department of Information Systems,
Eötvös Loránd University, Múzeum krt. 6–8, Budapest,
Hungary H-1068

evidence that the hippocampal formation is involved in relational processing (Eichenbaum et al. 1994; Young and Eichenbaum 1996)[1] the hippocampus encodes information about specific items and combinations of items if these items of information are to be held in the long-term memory. (2) The hippocampus encodes place fields and is responsible for spatial memory (see, e.g., O'Keefe and Nadel 1978; Burgess et al. 1995; Skaggs et al. 1996 and references therein). There is compelling evidence in favour of this twofold function of the hippocampus in the mammalian brain; the intriguing question is whether there is an unifying computational role that could include both.

Studies on adaptive goal-oriented systems utilizing episodic memories and reinforcement learning (Kalmár et al. 1995) led to the conclusion that generalizing concepts may be formed by means of an appropriate representation. Concepts were formed by peeling off (neglecting) those components of the representation that had no information about whether a given episode can happen or not. That is, a representation that minimizes mutual information between its components is preferable for generalization. Algorithms that minimize mutual information between representation components perform 'independent component analysis' and attempt to develop 'independent components'.

The present model makes use of the following assumption: long-term memory formation attempts do develop statistically independent memory traces, or independent components (IC). This assumption has two important consequences. (i) Independent components are not orthonormal and form a poor internal representation in feedforward networks. (ii) Independent components should be formed by some particular structure.

Independent component analysis, or blind source separation, has been the subject of extensive research starting with the original work of Jutten and Herault (1991). A basic algorithmic derivation was given by Comon (1994), and a connectionist solution was found by Bell and Sejnowski (1995). These developments gave rise to novel algorithms (Amari et al. 1996; Karhunen et al. 1997 and others). Here, architectures and learning rules developed by Oja and co-workers will be utilized because they fit the constraints of the model in a smooth fashion.

The model makes extensive use of a novel control architecture called the static and dynamic state (SDS) feedback control scheme (Szepesvári and Lőrincz 1996, 1997; Szepesvári et al. 1997). The model is made explicit by using a derivative of this control scheme, the data compression and reconstruction (DCR) architecture that utilizes relaxation dynamics (Fomin et al. 1997; Lőrincz 1997b). DCR architecture is closely related to the relaxation equations of Olshausen and Field (1996) that optimize information transfer. The Kalman filter approach of Rao and Ballard (1997) can be seen as a DCR architecture extended by predictive connections. At first sight, reconstruction dynamics is disadvantageous since it gives rise to delays. Realistic models of reconstruction dynamics should also take losses, i.e., the leaky nature of neural processing, into account, and this makes the possibility of reconstruction questionable.

Reconstruction dynamics may have an advantage over feedforward architectures. It offers a procedure to reach simultaneity at all levels since compensating the delays of the reconstructed input (locking the reconstructed input to the actual input) will simply eliminate all the delays of the internal representation. The suggested tool is the SDS control architecture. It will be shown that the composite DCR and SDS architecture develops new independent components, promotes simultaneity with the developing new components and also trains the architecture to hold the new independent components. The overall simultaneity of the computational units and the independent components allow the discoveries of causal relations, i.e., the discoveries of temporally ordered, higher order correlations that could be temporally connected via delayed associations. The higher order correlations can be made independent again, and the system can keep growing.

In short, the model assumes an SDS-like control architecture made of DCR-like layers with the control architecture playing a dual role: it performs temporal compensation, or locking, and it forms independent components. It will be shown that these assumptions smoothly fit the architectural constraints of the trisynaptic pathway of the hippocampus. The resulting framework justifies both concepts, i.e., the formation of place fields and the concept of relational processing.

I present here a functional model of the hippocampal-entorhinal loop. The model describes the activities of the biological substrates and the synaptic properties in artificial neural network terms such as activity vectors and memory matrices, respectively.

The paper is organized as follows. First (Sect. 2), a short description of the trisynaptic loop of the hippocampus is presented together with a description of some of the features of synaptic modifications in these areas. Then the two main building blocks, viz. the SDS control architecture (Sect. 3.2) and the DCR reconstruction architecture (Sect. 3.4), are reviewed. Section 4 derives a model that has two phases; the first phase corresponds to exploration when the architecture temporally compensates, while the second phase corresponds to memory consolidation when the architecture trains the long-term memory structures. In Sect. 5 a functional model of the hippocampal loop is formulated. Discussions about special aspects of hippocampal computations are provided in Sect. 6. Conclusions regarding the functions of the hippocampus are drawn in the last section. A short version of some of the ideas presented here has already been published (Lőrincz 1997a).

2 Some properties of the trisynaptic loop of the hippocampal formation

The superficial layers of the entorhinal cortex receive sensory information from the cortex. The same layers

[1] The importance of relational processing has been stressed by Taylor (1991 and manuscript submitted).

provide input to the hippocampal formation comprising three main subfields, the dentate gyrus and the CA3 and CA1 regions. The hippocampus processes this information and returns it to the deep layers of the entorhinal cortex. The same deep layers give rise to the efferents of the entorhinal cortex that return information to cortical targets where long-term memory traces are formed. These deep layers also provide inputs to the superficial layers of the entorhinal cortex (see, e.g., Iijima et al. 1996, and references therein). Figure 1 depicts the connection structure.

The trisynaptic loop of the hippocampus comprises three types of excitatory synapses. (1) The majority of entorhinal afferents form the perforant path and synapse onto granule cells of the dentate gyrus. (2) The main axons of the granule cells are the mossy fibre axons that form the characteristic mossy terminals on the thorny excrescences of the pyramidal cells of the CA3 subfield. Direct entorhinal afferents also reach the CA3 subfield. (3) A branch of the axons of the pyramidal neurons of the CA3 subfield that form the Schaffer collaterals innervate the apical dendrites of the principal cells of the CA1 subfield. The CA1 subfield represents the last stage of the intrahippocampal trisynaptic loop and is considered as the major output of the hippocampus. The CA1 subfield projects back to the entorhinal cortex both directly and via the subiculum.

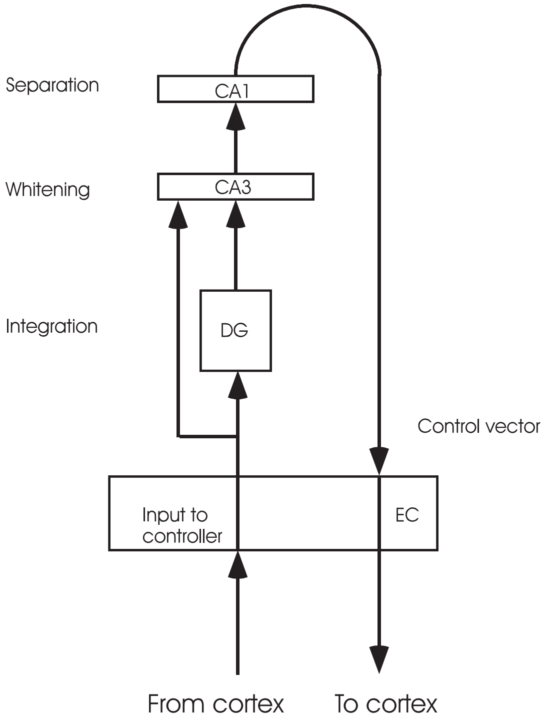

Fig. 1. Architecture of the temporal locking control scheme is made up of four networks. The first one, the entorhinal cortex (EC), is a reconstruction network receiving inputs from the cortex. The temporal locking architecture receives inputs computed at the input layer of this reconstruction network, while the output of the control scheme targets the output layer of the reconstruction network. The controller's input propagates via two independent channels, the first of which directly excites the CA3 subfield, whereas the other channel is temporally integrated at the dentate gyrus (DG) and then excites the CA3 subfield. The CA3 subfield is a whitening stage that decorrelates the inputs and gives rise to outputs of equal variances by means of the recurrent collaterals. The output of the CA3 subfield is the input of the CA1 subfield, which is a separation stage. (The CA3 and the CA1 networks also form reconstruction networks.) The output of the CA3 subfield serves a double purpose: it locks the input and the reconstructed input of the reconstruction network of the entorhinal cortex together and also trains the entorhinal cortex (and possibly other cortical areas) to hold the estimates of the independent components

As far as the intralayer connectivity is concerned, the CA3 subfield exhibits extensive commissural and associational collaterals, called recurrent collaterals. In contrast, the CA1 subfield exhibits limited interconnectivity (Christian and Dudek 1988).

The learning properties of the hippocampal formation are further complicated as can be seen from studies related to the different electroencephalogram (EEG) patterns of this area. Two of the EEG patterns are the theta waves (4–8 Hz) and the sharp waves of high-frequency (200 Hz) oscillations. In the CA1 subfield of the hippocampus of rats, sharp EEG waves can replace the regular 4–8 Hz theta waves during eating, drinking, sitting quietly, or slow-wave sleep (Freund and Buzsáki 1996 and references therein). According to the two-stage learning model of Buzsáki (1989), the sharp waves consolidate learning at synapses that were activated during exploratory behaviour, i.e. during the time of theta rhythm.

3 Building blocks of the model

Two related algorithms, a control algorithm based on speed-field tracking and a data compression and reconstruction architecture, are reviewed here.

3.1 Terminology

In this section the terms of the modeling are defined. Our artificial neural network model is built from distinct computational layers. Each computational layer performs one or more computational steps on all of the inputs it receives. The input of the layer can be the afferent input, which is typically some processed input of another computational layer. The output of the computational layer can serve as its own input, i.e., return loops can exist within the layered structure. The computation in each layer is performed in parallel by computational units, which make up the layer.

Let us try to connect artificial neural networks to real neurons by relating the firing rates of the neurons and activities of the artificial computational units. We consider that the average firing rate of a real neuron corresponds to zero activity for the artificial computational unit. Firing rates that are higher and lower than average correspond to positive and negative activities, respectively. The activity of a computational unit depends on the sum of the filtered inputs received through its input connections. In mathematical terms, such constructs can be described as follows. Assume that we have an input vector $\mathbf{x}$ and that $\mathbf{x}$ has N components, $\mathbf{x} = (x_1, \ldots, x_N)$ that is, $\mathbf{x} \in \mathbf{R}^N$. There are n computational units and the jth unit is denoted by u_j. In most cases it is assumed that the number of computational units is smaller or equal to the number of inputs, i.e., $n \leq N$. For the sake of notational simplicity, unit u_j has connections to each component of the input vector, but connections may have zero connection strengths. Let us denote the strength of the (feedforward) connection between the ith component of the input vector and unit u_j by q_{ij}. If q_{ij} is non-zero, then the input component i and computational unit u_j are connected synaptically. The components of the connection strength, belonging to the jth unit, can be arranged in vector form $\mathbf{q}_j^T = (q_{1j}, \ldots, q_{Nj})$ where superscript T denotes transposition. This vector is called the memory vector or connection strength vector to unit u_j. The memory vectors can be arranged in memory matrices. For example, memory vectors $\mathbf{q}_i$, $i = (1, \ldots, n)$ can be arranged to form memory matrix Q: the ith element of memory vector $\mathbf{q}_j$, i.e., q_{ij} is the ijth element of matrix Q. The activity of the jth unit is denoted by a_j. The activities of the computational units form the activity vector $\mathbf{a}$ ($\mathbf{a} \in \mathbf{R}^n$). The activity vector is sometimes called the internal representation of input $\mathbf{x}$. The activity vectors can form (part of) the inputs of computational layers. In what follows, vectors represent neural activities of layers, while matrices represent memory matrices between neural vectors.

3.2 Control architecture utilizing static and dynamic feedback

The control architecture performs speed-field tracking. Let $D \subseteq \mathbf{R}^N$ denote the domain of the plant's state with the equation of motion given by

$$\mathbf{u} = A(\mathbf{x})\underline{\mathbf{x}} + \mathbf{b}(\mathbf{x}) \tag{1}$$

where $\mathbf{x}$ is the state vector of the plant, the dot denotes temporal derivation, $\mathbf{u} \in \mathbf{R}^n$ is the control. Let us now assume that we have an estimate of the true inverse-dynamics function $\mathbf{\Phi}(\mathbf{x}, \dot{\mathbf{x}}) = A(\mathbf{x})\dot{\mathbf{x}} + \mathbf{b}(\mathbf{x})$, given by $\hat{\mathbf{\Phi}}(\mathbf{x}, \dot{\mathbf{x}}) = \hat{A}(\mathbf{x})\dot{\mathbf{x}} + \hat{\mathbf{b}}(\mathbf{x})$. The SDS feedback control equations can then be written as

$$\mathbf{u} = \mathbf{u}_f(\mathbf{x}, \dot{\mathbf{x}}, \mathbf{v}(\mathbf{x})) + \mathbf{w} \tag{2}$$

$$\dot{\mathbf{w}} = \Lambda\big(\hat{\mathbf{\Phi}}(\mathbf{x}, \mathbf{v}(\mathbf{x})) - \hat{\mathbf{\Phi}}(\mathbf{x}, \dot{\mathbf{x}})\big) \tag{3}$$

where $\mathbf{u}_f(\mathbf{x}, \dot{\mathbf{x}}, \mathbf{v}(\mathbf{x}))$ is the so-called feedforward controller, $\mathbf{w}$ is the so-called feedback controller, $\Lambda > 0$ is the gain of feedback, and the desired motion is determined by a speed-field tracking task that prescribes the desired speed vector $\dot{\mathbf{x}}$ of the plant as a function of the state vector:

$$\dot{\mathbf{x}} = \mathbf{v}(\mathbf{x}) \tag{4}$$

We assume that

$$\mathbf{u}_f(\mathbf{x}, \dot{\mathbf{x}}, \mathbf{v}(\mathbf{x})) = \hat{\mathbf{\Phi}}(\mathbf{x}, \mathbf{v}(\mathbf{x})) \tag{5}$$

It can be shown that under suitable conditions the scheme is uniformly ultimately bounded, that is, for all $\epsilon > 0$ there exists a gain Λ and absorption time $T > 0$ such that for bounded initial error ($||\mathbf{e}(0)|| < K\Lambda$) it holds that $||\mathbf{e}(t)|| < \epsilon$, provided $t > T$, and the solution can be extended up to time t. Here K is a fixed positive constant and $\mathbf{e}(0)$ denotes the initial value of $\mathbf{e}(t) = \mathbf{v}(\mathbf{x}(t)) - \dot{\mathbf{x}}(t)$, the state error. The proof is based on an extension of Liapunov's second method (Szepesvári et al. 1997). It can be seen that the 'feedforward controller' is actually a static state feedback controller in this scheme, justifying the shorthand, SDS scheme. The term 'feedforward' is nevertheless kept, since it describes a natural route of further generalizations.

Speed-field tracking is not typical in the control literature, but arises naturally if we consider stationary optimal-control problems such as path planning tasks (Hwang and Ahuja 1992). Conventional control tasks, such as *point-to-point control* and *trajectory tracking* cannot be exactly rewritten in the form of speed-field tracking and vice versa. The speed-field tracking task has the advantage that the designer can incorporate several objectives into the form of the speed-field to be tracked and hence extend the model's range of possibilities. This property will be exploited when locking the input and the reconstructed input.

This scheme can be interpreted as follows. We have an actual state ($\mathbf{x}$) and a desired speed ($\mathbf{v}$). We are equipped with the estimate of the inverse dynamics that provides us with a crude estimate of the true control vector ($\hat{\mathbf{\Phi}}(\mathbf{x}, \mathbf{v})$) that we term 'desired control vector'. We cannot use the desired control vector directly, since it is well-known that imprecise models of the inverse dynamics lead to instabilities. We utilize this control vector *together with* a correcting control vector. The correcting control vector is derived by measuring the actual speed and 'asking' the inverse dynamics 'What if the actual speed were the speed that we desire?' The estimate of the inverse dynamics provides us with the 'experienced control vector' ($\hat{\mathbf{\Phi}}(\mathbf{x}, \dot{\mathbf{x}})$) when input with the actual state ($\mathbf{x}$) and the actual speed ($\dot{\mathbf{x}}$). The difference between the desired and actual control vectors is then time integrated and used as the correcting means. The full scheme becomes a precise model of the inverse dynamics.

This control scheme is important as its working is almost identical to the reconstruction scheme to be developed later. The scheme can deal with first-order plants. Details can be found elsewhere (Szepesvári et al. 1997). Another solution emerges if the feedforward controller is replaced by the feedback controller.

$$\mathbf{u}_f(\mathbf{x}, \dot{\mathbf{x}}, \mathbf{v}(\mathbf{x})) = \hat{\mathbf{\Phi}}(\mathbf{x}, \mathbf{v}(\mathbf{x})) - \hat{\mathbf{\Phi}}(\mathbf{x}, \dot{\mathbf{x}}) \tag{6}$$

This scheme will be called the differencing control scheme. The differencing scheme exhibits global stability and can deal with plants of any order under suitable conditions (Szepesvári and Lőrincz 1997).

We note that temporal integration may include a temporal kernel of finite memory, and the stability theorem still holds. We also note that the actual state vector of the plant includes all of the variables that the equation of motion requires. Considering a rigid mechanical architecture, for example, the state vector of the plant includes both the positions of the masses as well as their momenta (Szepesvári and Lőrincz 1997).

3.3 Control architecture in case of fast error compensation

The differencing control architecture can be simplified under the assumption that error compensation is fast. In this case the temporal kernel can be of short duration, or of short temporal memory, and $\hat{A}$ can be pulled out from the temporal integration:

$$\mathbf{u} = \hat{A}(\mathbf{x})\left(\mathbf{v}(\mathbf{x}) - \dot{\mathbf{x}} + \Lambda \int K(t,t')(\mathbf{v}(t') - \dot{\mathbf{x}}(t'))dt'\right) \tag{7}$$

where $K(t,t')$ denotes the temporal kernel. Equation (7) has a term $(\mathbf{v}(\mathbf{x}) - \dot{\mathbf{x}})$ which is proportional to the error. The other term of (7) describes the temporally integrated error. Given that the state vector $(\mathbf{x})$ may contain the temporal derivatives also, temporal derivatives of the error are also compensated, and thus the architecture describes a generalized proportional, integral and derivative (PID) control architecture.

We note that these control architectures can compensate error within a subspace of the input space if the dimension of the control vector is smaller than the dimension of the input vector. Results on stability properties hold only for this subspace and may be further limited by the rank of matrix $\hat{A}(\mathbf{x})$.

3.4 Reconstruction architecture

The philosophy of the first-order control scheme allows one to derive a reconstruction architecture (Fomin et al. 1997; Lőrincz 1997b) equivalent to Wittmeyer's iterative scheme for solving matrix equations (Wittmeyer 1936). Assume that we have an input vector $\mathbf{x}$ ($\mathbf{x} \in \mathbf{R}^N$) and a 'direct internal representation' $\mathbf{a}_d$ ($\mathbf{a}_d \in \mathbf{R}^n$) connected by memory matrix Q ($Q \in \mathbf{R}^n \times \mathbf{R}^N$) formed by n memory vectors $\mathbf{q}_j$, ($j = 1, \ldots, n$) of dimension N. The relation between the input vector and the 'direct internal representation' is as follows:

$$\mathbf{a}_d = Q^T\mathbf{x} \tag{8}$$

The low-dimensional internal representation is then used to reconstruct the input by linearly superimposing the memory traces according to the internal representation. However, the direct internal representation needs correction to optimize reconstruction, and the correcting term is computed by means of a relaxation equation. The reconstructed input $\mathbf{y}$ ($\mathbf{y} \in \mathbf{R}^n$) is computed by means of the internal representation $\mathbf{a}$ and the memory matrix Q: $\mathbf{y} = Q\mathbf{a}$. The reconstructed input is then compared to the original input, and the difference is evaluated (filtered) by the memory matrix, giving rise to a correcting term to the internal representation, just like the correction vector served to develop the control vector of high precision. Relaxation of the internal representation $\mathbf{a}$ stops when the internal representation of the input and that of the reconstructed input become identical (Fomin et al. 1997; Lőrincz 1997b).

$$\mathbf{a} = \mathbf{a}_d + \mathbf{w} \tag{9}$$

$$\dot{\mathbf{w}} = \lambda Q^T(\mathbf{x} - \mathbf{y}) \tag{10}$$

These two equations can be collected into one equation that contains temporal integration

$$\mathbf{a} = \mathbf{a}_d + \lambda \int Q^T(\mathbf{x} - \mathbf{y})dt \tag{11}$$

The architecture can assume different forms (Lőrincz 1997b). The scheme can be extended by leaky integration. One form of introducing leaky integration is as follows:

$$\mathbf{a} = \mathbf{a}_d + Q^T\mathbf{c} \tag{12}$$

$$\mathbf{y} = Q\mathbf{a} \tag{13}$$

$$\dot{\mathbf{c}}_x = -\mathbf{c}_x + \lambda\mathbf{x} \tag{14}$$

$$\dot{\mathbf{c}}_y = -\mathbf{c}_y + \lambda\mathbf{y} \tag{15}$$

$$\mathbf{c} = \mathbf{c}_x - \mathbf{c}_y \tag{16}$$

where $Q^T\dot{\mathbf{c}}$ is equal to $\dot{\mathbf{a}}$ provided that $\mathbf{x}$ is constant. The corresponding architecture is depicted in Fig. 2. The boxes with integration sign represent leaky integration. Leaky integration can be viewed as synaptic integration (Markram and Tsodyks 1997). Leaky integration limits the reconstruction capabilities. In (11) and (12) the term $\mathbf{a}_d$ can be left out. This term is important if the memory vectors are orthonormal because then the network becomes a fast feedforward architecture.

The leaky equations can be rewritten in the form of the relaxation equation of Olshausen and Field (1996) that optimizes information transfer and representation sparseness simultaneously provided that the dimension

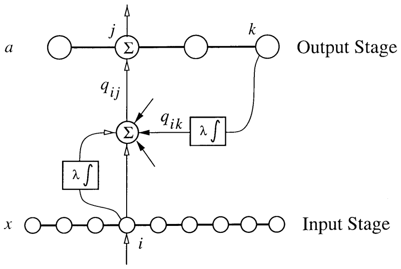

Fig. 2. Architecture of the data compression and reconstruction (DCR) scheme. The input gives rise to direct and integrated excitations of output, and, via reconstruction, an integrated inhibition of the output. The integrations are both leaky and together form the correcting term of the scheme. (The reconstructed input is not explicitly represented by any of the nodes of the figure)

of the internal representation is equal or larger than the dimension of the input and that a non-linear leaky term is introduced. In (14) the dimension of the internal representation is smaller or equal to the dimension of the input, and if the equation were not leaky, then upon relaxation the equation

$$Q^T\mathbf{x} = Q^T Q\mathbf{a} \tag{17}$$

would hold. This equation corresponds to Wittmeyer's iterative algorithm (Wittmeyer 1936) and solves the equation $\mathbf{x} = Q\mathbf{a}$ for $\mathbf{a}$ under the condition that the mean square error $\|\mathbf{x} - Q\mathbf{a}\|^2$ is to be minimized. For notational simplicity, in the followings (12)–(16) will be condensed into the two equations

$$\mathbf{a} = \mathbf{a}_d + \mathbf{w} \tag{18}$$

$$\dot{\mathbf{w}} = -\mathbf{w} + \lambda Q^T(\mathbf{x} - Q\mathbf{a}) \tag{19}$$

We note that reconstruction networks have a large degree of freedom. The most important part of the network is the reconstruction equation, $\mathbf{y} = Q\mathbf{a}$. The feedforward equations, however, may contain different matrices, e.g., one may replace matrix Q^T by P^T, use $\mathbf{a} = P^T(\mathbf{x} + \mathbf{c})$, provided that matrices Q^T and P^T span the same subspace of $\mathbf{R}^N$.

4 Control model for error compensation and independent component formation

4.1 Error compensation for temporal locking

Armed with the SDS and DCR schemes we are in a position to develop a unifying model for the formation and the usage of independent components. Independent components may strongly overlap, and thus the columns of the memory matrix are not orthonormal. Then the internal representation becomes poor unless some means, e.g. reconstruction networks, are utilized. In the case of feedback architectures such as the reconstruction networks, however, reconstruction will be delayed, and temporal compensation becomes an important issue. The leaky nature of (14) and (15) represents another obstacle to reconstruction.

The key to escaping this trap is to force reconstruction. The reconstructed input should somehow be upgraded to match the actual input since then there will be no delay in the internal representation, and the afore said drawback disappears. In fact, this becomes an advantage over feedforward processing since matching at the input layer by means of the output activities can approximate instantaneous processing.

Let us formulate the problem in terms of our error compensating control architecture. Our desired reconstruction vector is $\mathbf{x}$, while our experienced reconstruction vector is $\mathbf{y}$. Assuming that error compensation is fast, we can use the generalized PID architecture, and the control vector may be written as

$$\mathbf{u}(\mathbf{x}, \mathbf{x} - \mathbf{y}) = \hat{A}(\mathbf{x})\left(\mathbf{x} - \mathbf{y} + \Lambda \int K(t, t')(\mathbf{x}(t') - \mathbf{y}(t'))dt'\right) \tag{20}$$

where control vector $\mathbf{u}$ is considered as the internal representation that reconstructs input $\mathbf{x}$ by means of memory matrix Q, that is, $\mathbf{a} = \mathbf{u}$ and

$$\mathbf{y} = Q\mathbf{u}(\mathbf{x}, \mathbf{x} - \mathbf{y}) \tag{21}$$

where Q is the long term memory matrix of the model which represents the long-term memory of the entorhinal cortex. According to the differencing SDS scheme we obtain the desired result that the reconstruction error becomes small, and thus the control architecture makes reconstruction feasible.

The architecture can be described as follows (see Fig. 1). There is a computational layer (tentatively called the entorhinal cortex) that receives processed sensory information from other networks. This input undergoes a reconstruction procedure by means of a loop structure. The input to the loop is the difference between the input to the entorhinal cortex and the reconstructed input. The reconstructed input is formed by the output of the loop structure and the memory matrix of the entorhinal cortex. The loop structure forms a generalized PID controller with temporally integrated and direct error channels that are followed by an associative stage. (The associative stage consists of two layers in Fig. 1. These layers are tentatively denoted by CA3 and CA1.) The generalized PID controller barely restricts the specific properties of the associative stage since the only constraint against this stage is the 'sign-proper' feedback of the error. It is possible, for example, that matrix $\hat{A}(\mathbf{x})$ is a highly non-linear function of its input arguments. As long as the sign of the feedback is correct, the controller can work with this non-linearity at the cost that it may affect the absorption time T of the controller and the magnitude of the reconstruction error $\|\mathbf{e}(t)\|$.

The associative stage will be derived in the following subsections. In the derivation we assume that the associative stage has another role; it develops independent components.

4.2 Associative stage

The requirement of temporal locking led to the need of an associative stage in a natural fashion. We require that the associative stage develops statistically independent, i.e. separated, outputs (Jutten and Herault 1991; Comon 1994; Bell and Sejnowski 1995; Wang et al. 1995).

In networks dealing with independent component analysis (ICA), it is assumed that signals are provided by independent sources: in other words, the joint probability density of the components of the original signal vector $\mathbf{s} = (s_1, \ldots, s_k)$ can be given as the product of the marginal densities of the individual components. These signals are mixed and are covered with noise, i.e. the input to the system is

$$\mathbf{x} = R^T\mathbf{s} + \mathbf{n} \tag{22}$$

where $\mathbf{n}$ denotes the noise, R is a constant mixing matrix with full column rank, and the presentation index (time) has been dropped for simplicity. The task is to develop memory traces so that components of the neural activity vector represent components of vector $\mathbf{s}$.

Association that leads to independent components (IC) can be built up from three stages (Wang et al. 1995; Karhunen et al. 1997). There is a first stage that whitens, i.e., decorrelates and rescales the input, giving rise to outputs of equal variance. The second stage makes use of the whitened information and does 'blind source separation' (Jutten and Herault 1991; Comon 1994; Bell and Sejnowski 1995; Wang et al. 1995). Upon training, the outputs of this stage become statistically independent. The outputs of the second stage are then used at a third stage to train a memory matrix. The columns of that matrix represent the individual independent components.

According to the anatomical constraints these three stages correspond to the CA3 and the CA1 subfields of the hippocampal formation and the entorhinal cortex itself, with the entorhinal cortex holding the ICs. The corresponding processing layers will tentatively be labelled by CA3, CA1 and EC, repectively. In what follows, reconstruction architectures will be considered the substrates of whitening and separation, as well as the storage place of long-term memories. It should be noted that the utilization of reconstruction architectures at every level is not a necessity. It is only the coherence of the model that makes this choice appealing. Another note to make is that the preparation of the input for the controller (i.e. differencing) does not spoil separation since it corresponds to mixing provided that the inputs are temporally adjusted and locked. This condition is consistent with the setting of the model.

4.2.1 Whitening with reconstruction architecture

Whitening in the reconstruction network can be achieved by introducing recurrent collaterals. Denoting the matrix representing the recurrent collaterals by O_{CA3}, (14) can be modified accordingly, and the activity vector $\mathbf{a}_{\mathrm{CA3}}$ can be written as follows:

$$\mathbf{a}_{\mathrm{CA3}} = \mathbf{a}_{d,\mathrm{CA3}} + \mathbf{w}_{\mathrm{CA3}} \tag{23}$$

where $\mathbf{a}_{d,\mathrm{CA3}} = Q^T_{\mathrm{CA3}}\mathbf{x}_{\mathrm{CA3}}$ and

$$\begin{aligned}\dot{\mathbf{w}}_{\mathrm{CA3}} = & -\mathbf{w}_{\mathrm{CA3}} + \lambda_1 Q^T_{\mathrm{CA3}}(\mathbf{x}_{\mathrm{CA3}} - \mathbf{y}_{\mathrm{CA3}}) \\ & + \lambda_2 Q^T_{\mathrm{CA3}} O_{\mathrm{CA3}}\mathbf{a}_{\mathrm{CA3}}\end{aligned} \tag{24}$$

where

$$\mathbf{y}_{\mathrm{CA3}} = Q_{\mathrm{CA3}}\mathbf{a}_{\mathrm{CA3}} \tag{25}$$

We set $\lambda_1 = \lambda_2 = \lambda$. We assume identical training rules for matrices Q_{CA3} and O_{CA3} because elements of identical indices of these matrices connect the same input and output nodes of the network. Thus, upon training, the two matrices become equal. Then, in turn, the recurrent collaterals cancel the effect of the reconstructing term $Q_{\mathrm{CA3}}\mathbf{y}_{\mathrm{CA3}}$. If the recurrent collaterals are in effect, then

$$\mathbf{a}_{\mathrm{CA3}} = (1+\lambda)Q^T_{\mathrm{CA3}}\mathbf{x}_{\mathrm{CA3}} \tag{26}$$

We suggest the following training rules:

$$\Delta Q_{\mathrm{CA3}} = \eta_1(Q_{\mathrm{CA3}} - \eta_2\mathbf{y}_{\mathrm{CA3}}\mathbf{a}^T_{\mathrm{CA3}}) \tag{27}$$

and a similar equation for matrix O_{CA3}:

$$\Delta O_{\mathrm{CA3}} = \eta_1(Q_{\mathrm{CA3}} - \eta_2\mathbf{y}_{\mathrm{CA3}}\mathbf{a}^T_{\mathrm{CA3}}) \tag{28}$$

These learning equations assume relatively high firing rates and can be justified as follows. Hebbian learning requires two units to fire in a 'coincident' manner. Exact coincidence is not required, coincidence means 'within a narrow temporal window'. At high firing rates, input and output may fire within this window of coincidence for unrelated patterns. Then coincidence becomes a statistical property; the stronger the synapse, the more it contributes to firing and the more probable that within this narrow temporal window unrelated 'coincidences' are experienced by the synapse. Synapses become 'self-strengthening' at high firing rates. This self-strengthening property is represented by the first term of (27) and (28). The self-strengthening effect is limited by a Hebbian term proportional to both the activity of the neuron and the recurrent inhibitory activity represented by the appropriate component of the reconstruction vector $\mathbf{y}_{\mathrm{CA3}}$.

When the recurrent collaterals are in effect, that is when (26) holds, the learning equations (27) and (28) can be written as

$$\begin{aligned}\Delta Q_{\mathrm{CA3}} &= \Delta O_{\mathrm{CA3}} \\ &= \eta_1 Q_{\mathrm{CA3}}(I - \eta_3 Q^T_{\mathrm{CA3}}\mathbf{x}_{\mathrm{CA3}}\mathbf{x}^T_{\mathrm{CA3}}Q_{\mathrm{CA3}})\end{aligned} \tag{29}$$

where I denotes the identity matrix and $\eta_3 = \eta_2(1+\lambda)^2$. We assume that η_1 is rate dependent and vanishes at low rates. Apart from the constant factor η_3, the learning equations (27, 28) are exactly the whitening equation of Laheld and Cardoso (1994). These learning equations form an orthogonal memory system with memories of equal norms. For the sake of simplicity we assume that $\eta_3 = 1$, in which case the learning equations form a projector. We have thus the following important consequence: if the recurrent collaterals are not in effect, then the internal representation can be written as

$$\mathbf{a}_{\mathrm{CA3}} = Q^T_{\mathrm{CA3}}\mathbf{x}_{\mathrm{CA3}} \tag{30}$$

because matrix $Q^T_{\mathrm{CA3}}Q_{\mathrm{CA3}}$ is the identity matrix. Thus, one can turn the recurrent collaterals on and off, and with training they will have the same outputs apart from a scaling factor for both cases. The scaling factor depends on the recurrent collaterals, whether they are on or off.

4.2.2 Separation with reconstruction architecture

The next step of the independent component analysis is a separation stage. Of the many possibilites we choose one model that fits the architectural constraints of the CA1 subfield such that the CA1 subfield exhibits limited within-layer excitatory interconnectivity, this being a contrasting difference between the CA3 and CA1 subfields (Christian and Dudek 1988). We shall build up the learning rule from two terms. We assume that the losses have different functional forms during the theta and sharp wave phases. The CA1 subfield is more active during the sharp wave phase than during the theta phase when the CA1 subfield is relatively quiet. We assume that during the sharp wave phase inhibitory terms including losses, and reconstructing effects are small. Then the activity vector may be approximated as

$$\mathbf{a}_{\mathrm{CA1}} = Q^{T}_{\mathrm{CA1}} \mathbf{x}_{\mathrm{CA1}} \tag{31}$$

Just like for subfield CA3, the statistical learning rule (27) is suggested to govern adaptation during the high firing rate sharp wave (SPW) phase:

$$\Delta Q_{\mathrm{CA1}} = \eta_{\mathrm{SPW}} Q_{\mathrm{CA1}} (I - \mathbf{a}_{\mathrm{CA1}} \mathbf{a}^{T}_{\mathrm{CA1}}) \tag{32}$$

where the relation $\mathbf{y}_{\mathrm{CA1}} = Q_{\mathrm{CA1}} \mathbf{a}_{\mathrm{CA1}}$ was taken into account. We assume that η_{SPW} is rate dependent and vanishes at low rates. In case of a tuned matrix Q_{CA1}, the expectation value of the left-hand side is equal to zero. This condition is satisfied when the expectation value of $\mathbf{a}_{\mathrm{CA1}} \mathbf{a}^{T}_{\mathrm{CA1}}$ becomes equal to the identity matrix. Let us note that the CA1 subfield receives input from whitened signals, that is, the expectation value of $\mathbf{x}_{\mathrm{CA1}} \mathbf{x}^{T}_{\mathrm{CA1}}$ is equal to the identity matrix. Then (32) is equivalent to the following tuning rule:

$$\Delta Q_{\mathrm{CA1}} = \eta_{\mathrm{SPW}} Q_{\mathrm{CA1}} (I - Q^{\mathrm{T}}_{\mathrm{CA1}} Q_{\mathrm{CA1}}) \tag{33}$$

During the relatively quiet theta phase we assume that non-linear losses limit the firing and that the dominant loss terms are connected to neural activities. Equations (18) and (19) are slightly modified to fit these assumptions:

$$\dot{\mathbf{a}}_{\mathrm{CA1}} = -G(\mathbf{a}_{\mathrm{CA1}}) + \lambda Q^{T}_{\mathrm{CA1}} (\mathbf{x}_{\mathrm{CA1}} - Q_{\mathrm{CA1}} \mathbf{a}_{\mathrm{CA1}}) \tag{34}$$

where $G(.)$ denotes the same sharply increasing non-linear function for all of the components. In the case of high non-linear losses, one may assume that the activities are small and that the term $Q_{\mathrm{CA1}} \mathbf{a}_{\mathrm{CA1}}$ on the right-hand side of (34) may be neglected. Then, in turn, $\mathbf{a}_{\mathrm{CA1}}$ may be approximated as:

$$\mathbf{a}_{\mathrm{CA1}} = g(Q^{T}_{\mathrm{CA1}} \mathbf{x}_{\mathrm{CA1}}) \tag{35}$$

where $g(.)$ denotes the inverse function of $\lambda^{-1} G(.)$. Function $g(.)$ can be a saturating function. One may consider, for example, the hyperbolic tangent function tanh(.) as a candidate for all of the components of function $g(.)$. We suggest that during the low firing rate theta phase, matrix Q_{CA1} undergoes Hebbian learning. According to Fig. 2, Hebbian learning occurs between the reconstruction error and the neural activity value. Thus, during the theta phase the learning rule may be given as follows:

$$\Delta Q_{\mathrm{CA1}} = \eta_{\mathrm{theta}} (\mathbf{x}_{\mathrm{CA1}} - Q_{\mathrm{CA1}} \mathbf{a}_{\mathrm{CA1}}) \mathbf{a}^{T}_{\mathrm{CA1}} \tag{36}$$

The term $Q_{\mathrm{CA1}} \mathbf{a}_{\mathrm{CA1}}$ can be neglected during this part of the training. Then we can summarize the two-phase learning rule of the CA1 subfield as follows:

$$\begin{aligned} \Delta Q_{\mathrm{CA1}} = {} & \eta_{\mathrm{SPW}} Q_{\mathrm{CA1}} (I - Q^{\mathrm{T}}_{\mathrm{CA1}} Q_{\mathrm{CA1}}) \\ & + \eta_{\mathrm{theta}} \mathbf{x}_{\mathrm{CA1}} g^{\mathrm{T}} (Q^{\mathrm{T}}_{\mathrm{CA1}} \mathbf{x}_{\mathrm{CA1}} \end{aligned} \tag{37}$$

This equation is the so-called bigradient separating algorithm working on whitened inputs (Wang et al. 1995; Karhunen et al. 1997) and thus the model CA1 subfield separates.

We note that non-linear operation is assumed during the theta phase and linear operation is assumed during the sharp wave phase. This means that the network provides separated outputs only during the sharp wave phase, whereas in the theta phase it works as a saturating controller.

4.2.3 Learning the independent components

The last stage of independent component analysis is the stage where the ICA basis vectors are to be estimated, i.e. learned. According to the anatomical constraints of the model, this stage is either the entorhinal cortex itself or the association cortices with entorhinal afferents. We shall tentatively use the subscript EC to describe the computational properties of this network. The learning rule of this stage can be given as (Karhunen et al. 1997):

$$\Delta Q_{\mathrm{EC}} = \eta (\mathbf{x}_{\mathrm{EC}} - Q \mathbf{a}_{\mathrm{CA1}}) \mathbf{a}^{T}_{\mathrm{EC}} \tag{38}$$

where vector $\mathbf{x}_{\mathrm{EC}}$, $\mathbf{a}_{\mathrm{EC}}$ and Q_{EC} describe the input, the internal representation and the memory matrix to the entorhinal cortex, respectively. According to the loop structure it is also true that

$$\mathbf{a}_{\mathrm{EC}} = \mathbf{u}_{\mathrm{loop}} = \mathbf{a}_{\mathrm{CA1}} \tag{39}$$

Expression $\mathbf{a}_{\mathrm{EC}}$ denotes the output of the controller, i.e. the output of the CA1 subfield in its linear (sharp wave) phase when the CA1 subfield provides separated outputs. This training rule minimizes the mean square error of expression $\|\mathbf{x}_{\mathrm{EC}} - Q_{\mathrm{EC}} \mathbf{a}_{\mathrm{CA1}}\|^2$ so that the rows of matrix Q_{EC} become independent components. When the reconstruction process is in effect, the full network including the control loop relaxes to vector $\mathbf{u}_{\mathrm{loop}}$ that minimizes the expression $\|\mathbf{x}_{\mathrm{EC}} - Q_{\mathrm{EC}} \mathbf{u}_{\mathrm{loop}}\|$ where the particular form of the norm depends on the non-linearity of the CA1 subfield. Thus, the reconstruction net always minimizes expression $\|\mathbf{x}_{\mathrm{EC}} - Q_{\mathrm{EC}} \mathbf{a}_{\mathrm{EC}}\|$, while the loop tunes the components of matrix Q_{EC} to form independent components; matrix Q_{EC} approximates matrix R^{T} of (22). The resulting architecture is depicted in Fig. 1.

5 Functional model of the hippocampal loop

We are in the position to develop a functional model of the entorhinal cortex and the hippocampal formation. First we suggest that the hippocampal formation and the entorhinal cortex together form a loop and that this loop corresponds to a dynamic autoassociator that renders a reconstruction vector to the input vector of the entorhinal cortex. The reconstruction process optimizes the internal representation, which would be poor if feedforward networks made of non-orthonormal memory vectors, such as independent components, were to be utilized.

It is the difference between the input vector and the reconstruction vector of the entorhinal cortex which enters the hippocampal loop. This difference can excite the CA3 subfield in a direct fashion. This difference also enters the dentate gyrus and forms, in turn, the dentate afferents of the CA3 subfield. The full loop can be viewed as a control architecture that resembles PID controllers. The direct entorhinal afferents of the CA3 subfield form the proportional term, while the dentate afferents of the CA1 subfield form the temporally integrated term. Integration can be accomplished by means of recurrent feedback excitations, and such constructs can be found in the dentate gyrus (Li et al. 1994). Derivative information may be conveyed to the CA3 subfield if such information is input into the entorhinal cortex.

The afferents of the CA3 subfield thus form error terms that undergo temporal operations, such as integration and derivation. During the theta phase the PID controller then outputs a non-linear control vector by means of an associative stage, the error-to-control-vector association. The PID controller requires sign-proper association, which leaves a large degree of freedom in the design of the associative stage. We propose that the associative stage is built to form independent components. We suggest that independent component formation is accomplished by means of a two-phase operation scheme. The first phase corresponds to exploration when the temporal compensation is important since temporal compensation speeds up the formation of the internal representation. The second phase corresponds to memory consolidation when the hippocampal loop provides separated linear output signals and those signals are used for memory formation in the entorhinal cortex. These suggestions are in agreement with others in the literature (Buzsáki 1989) that the theta phase corresponds to exploration, while the sharp waves correspond to consolidation of long-term memory.

The associative stage consists of two layers, a whitening stage and a separation stage. It is suggested that the whitening stage is a reconstruction layer with recurrent collaterals and that the CA3 subfield corresponds to this whitening stage. The training within this layer is such that – apart from a scaling factor – the output of the CA3 subfield is the same both when these collaterals are effective and also when they are not. Switching occurs between the theta and sharp wave phases; in the theta phase the recurrent collaterals are inhibited by subcortical inputs. We assume that the learning equations that correspond to the whitening equation of Laheld and Cardoso (1994) give rise to changes in the recurrent collaterals during the theta phase and that these changes rule the sharp wave bursts during the sharp wave phase. The feature that the output of the CA3 subfield – apart from a scaling factor – is the same during both phases is important because the same whitened output is needed both when training the separation stage and when training the storage place of the long-term memories.

Separation of the whitened signals is then performed by the CA1 subfield, which operates differently in the two phases. During the theta phase it is relatively quiet as it operates in its non-linear mode. During this theta phase the outputs of the CA1 subfield control the reconstruction process in the entorhinal cortex. Non-linear processing in the CA1 subfield does not prevent reconstruction in the entorhinal cortex as long as the non-linearity of the CA1 subfield does not change the sign of the output of this subfield, because the PID control architecture allows saturating non-linearities that do not modify the sign of the feedback (Szepesvári and Lőrincz 1997). Also, these non-linear outputs of the CA1 subfield tune the memory vectors of the CA1 subfield. The learning rule forms orthonormal and separating memory components within the CA1 subfield.

During the sharp wave phase the CA1 subfield operates in its linear mode. The outputs of the CA1 subfield during this linear mode then train the long-term memory system of the entorhinal cortex to form independent components.

5.1 Hebbian learning

The learning rule of the entorhinal cortex is of Hebbian type provided that the reconstruction error is represented somehow at the input layer during the sharp wave phase. This assumes that synapses of the entorhinal cortex undergo transient changes during the theta phase and that these transient changes are consolidated during the sharp wave phase. If so, then learning takes place in those synapses of the entorhinal cortex that connect active units which represent the control output of the hippocampal formation to active units which represent the reconstruction error. These synapses connect the deep layers of the entorhinal cortex to the superficial layers. The learning in the entorhinal cortex is supervised.

The learning rule of the CA3 subfield is Hebbian. The learning rule has no leaky term but rather incorporates a self-strengthening contribution. This self-strengthening contribution is then limited by the Hebbian term that has a negative sign in this particular case. Learning is unsupervised; the activities that govern Hebbian learning are computed by the dynamics of the layer.

According to the model the learning rule of the CA1 subfield is divided into two phases. In the high firing rate sharp wave phase, the learning rule is similar to that of the CA3 subfield. In the low firing rate theta phase, Hebbian learning making use of the reconstruction error of CA1 layer is assumed. Learning is unsupervised.

6 Discussion

Here, I shall discuss the role of the hippocampal formation in producing long-term declarative memory traces.

The present model of the hippocampal formation formulates how and why this area is responsible for relational processing (Eichenbaum et al. 1994; Young and Eichenbaum 1996). The model extends the original idea by claiming that the basic building blocks of relational processing are the independent components. Internal representations using feedforward architectures and memory vectors of independent components, however, form a rather poor representation owing to the fact that independent components may strongly overlap and thus most of the internal components are excited by most of the inputs. It is thus suggested that the entorhinal cortex and the hippocampal loop form a reconstruction network since the constraints raised by input reconstruction optimize the internal representation. The dynamic optimization process exhibits itself as apparent competition between the nodes forming internal representation (Fomin et al. 1997).

The model assumes that the hippocampal formation is a control architecture that promotes temporal compensation to speed up reconstruction and to compensate for losses in the processing. This gives rise to fast pseudoinverse computation. The output of the controller also serves as a separation signal that provides statistically independent outputs and can train long-term memory traces to hold such memories. The model does not assume that information is kept in the hippocampus forever, but rather that the hippocampus learns to separate data and then trains the long-term memory to hold independent components. It is suggested that the new independent components can be considered as new 'sensors' that can be called 'relational sensors'. By relational sensors we mean, e.g., the neurons that can be found along the superior temporal polysensory area (STPa) that fire for 'compatible' form and motion (e.g. left profile moving left) or others that fire for 'incompatible' form and motion (e.g. right profile moving left) (Oram and Perrett 1996). These relational sensors may be considered as sensors of approximately independent components. This system can keep growing: such sensors can be combined at another stage, and higher-order decorrelation and higher-order separation can take place, allowing eventually higher-order relational sensors to form. To give another example, let us consider cells that are sensitive to edges. It has been argued that edges are the independent components of natural scenes (Olshausen and Field 1996; Bell and Sejnowski 1997). One can thus say that an edge-sensitive cell fulfils our requirements for relational sensors and represents the spatial relations of some lower-order cells. At higher levels more complex relational sensors are formed from the edge sensors, such as sensors responding to particular shapes or faces, etc. (Tanaka et al. 1991; Logothetis et al. 1995). Since some of these cells can encode invariant properties, either information on different views or information on possible flow fields that allow dynamic remapping is coded by these cells. According to our assumptions these cells code statistically independent properties and that could be revealed if reconstruction was not in effect. Combining these shape sensors and motion sensors, new relational sensors (e.g. 'compatible' and 'incompatible' sensors) can be formed, and so on.

Making use of the assumption that long-term memory forms independent components, then upon training these independent components manifest themselves as high-order correlations. In the simplest case independent components would correspond to edges (Olshausen and Field 1996; Bell and Sejnowski 1997). Edges are thus overlapping higher correlations of natural scenes, and their independence is to be measured by taking into account the size of the memory that holds these components. Upon training more complex overlapping higher-order correlations appear and can be represented in a hierarchical manner by means of further processing stages. We assume that hippocampal cells represent such complex higher-order correlations of overlapping features describing quasi-independent regions of the environment. The result is that hippocampal cells will eventually represent place cells, the ultimate higher-order correlations in a labyrinth. In other words, a cell will fire robustly when the animal is in a specific small portion of the environment and will be virtually silent everywhere else [O'Keefe and Nadel 1978; Burgess et al. 1995; Skaggs et al. 1996).

The model claims that reconstruction errors give rise to independent component formation by means of the hippocampus. The model thus claims that the hippocampus processes novel information and builds up memory traces in the entorhinal cortex and in the neocortex that represent this novel information. This transfer of novel information, in turn, makes the hippocampus available for novel information, that is, for afferent inputs that the loop of the hippocampus and the entorhinal cortex cannot reconstruct.

The model is consistent with the experimental findings that the hippocampus is crucial in forming long-term memories. The model also says that memories can be held in the hippocampus and that memories in the entorhinal cortex are governed by the hippocampal output. Thus, a hippocampal lesion may also effect retrograde memories.

The model is a functional model. If a realistic model of the hippocampal formation is to be constructed, then it should be mapped onto the biological substrate. Such mapping should take into account several important points of hippocampal formation including phase coding properties (O'Keefe and Recce 1993; Skaggs et al. 1996), features of the inhibitory networks, the topographical mapping between different areas, the different entorhinal afferents of the CA3 and the CA1 subfields, etc. We note first that the topographical mapping between the entorhinal cortex and the CA1 subfield is a straightforward consequence of the model that identifies the outputs of the CA1 subfield and the internal representation of the entorhinal cortex. Other properties such as phase coding, however, may not be included into the model in a straightforward fashion. The model is attractive from the following points of view: (i) several properties of the hippocampal-entorhinal loop are natural consequences of the assumption that long-term memories are made of independent components; (ii) the generalized PID architecture and the reconstruction networks are both flexible concepts and are thus promising candidates to form a biological model.

7 Conclusions

A speed-field tracking control architecture comprising layers featuring reconstruction dynamics has been suggested to describe the computational tasks of the hippocampal formation. According to the model, hippocampal formation is responsible for the temporal compensation of the various information processing units in the entorhinal cortex and for the process by which independent components are formed and are consolidated in the long-term memory. It is suggested that independent components can be considered as new relational sensors. The underlying reason for an additional architecture that consolidates long-term memory is the need for the development of independent components.

Independent components need reconstruction networks since without the constraint of reconstruction, internal representation developed by the non-orthogonal independent components can be poor. Reconstruction networks raise another problem, however, in that for highly overlapping inputs, reconstruction becomes slow. Also, owing to leaky integration, reconstruction can be still poor. The suggested control architecture that solves these problems, features (1) speed-field tracking, (2) robust control and (3) arbitrarily small error (Szepesvári and Lőrincz 1996, 1997; Szepesvári et al. 1997).

It was shown that reconstruction architectures are ideally suited for temporal compensation owing to the fact that the input and the reconstructed input can be locked together and that this condition promotes simultaneity at the level of the internal representation. The requirement that long-term memory should hold independent components set architectural constraints for the model. These constraints smoothly fit the anatomical features of the hippocampal formation. In particular, the loop structure, the major stages (the dentate gyrus, the CA3 and the CA1 subfields), the interconnectivity of these areas (the so-called trisynaptic pathway that includes direct and indirect excitations of the CA3 subfield, and the intralayer interconnectivity of the CA3 and the CA1 subfields) as well as the input-output connection structure of the entorhinal cortex are described by the model. According to the model the dentate gyrus is a simple integration stage. The CA3 subfield forms a whitening stage, while the CA1 subfield provides a separation stage with independent outputs. The nonlinearly processed output of the hippocampal formation is the control signal that locks the input and the reconstructed input of the entorhinal cortex to each other and also the signal that forms the memory components of the CA1 subfield via Hebbian learning. At the same time the linearly processed output of the hippocampal formation is precisely the training signal that can develop the independent components of the entorhinal cortex via Hebbian learning.

The model thus requires a two-phase operation. In the first phase the input and the reconstructed input of the entorhinal cortex are locked to each other, and thus the internal representation is not delayed. We identify this phase with the theta phase and thus with exploratory behaviour. The second phase is the memory consolidating phase when the output of the CA1 subfield trains the long-term memories to hold independent components. This phase is identified with the sharp wave phase. The model thus agrees with the original suggestion of Buzsáki (1989).

The model suggests that the entorhinal cortex and the neocortex become the storage place of the independent components. Also, the memories held by the entorhinal cortex are ruled by the hippocampal outputs. Novel memories are held by the hippocampus itself. Thus, according to the model a hippocampal lesion leads to two effects: (1) the formation of long-term memories deteriorates and (2) parts of the long-term memories are lost.

The model is a functional model, and several points require further investigation to validate it, including the role of subcortical inputs, the topographical order between layers of the hippocampal-entorhinal loop, the inhibitory networks of the hippocampal loop, the phase coding properties of these areas, etc. The model is promising from the point of view that the basic assumption that long-term memories are made of independent components fits smoothly with both the related control architecture and the reconstruction network and that these architectures may be flexible enough to meet the constraints of the biological substrate.

Acknowledgements. Helpful discussions with Péter Aszalós, László Balázs, Gyula Kovács, and Csaba Szepesvári are gratefully acknowledged. Special thanks are due to György Buzsáki for helpful comments on the manuscript. This work was partially supported by OTKA (Hungary) grants T14566 and T17100 and by grant JF no. 519 of the US-Hungarian Joint Fund.

References

Abraham WC, Gustafsson B, Wigstrom H (1986) Single high strength afferent volleys can produce long-term potentiation in the hippocampus in vitro. Neurosci Lett 70:217–222

Amari SL, Cichoki A, Yang HH (1996) A new learning algorithm for blind signals. In: Advances in Neural Information Processing Systems 8. MIT Press, Cambridge, Mass

Bell AJ, Sejnowski TJ (1995) An information-maximization approach to blind separation and blind deconvolution. Neural Comput 7:1129–1159

Bell AJ, Sejnowski TJ (1997) Edges are the independent components of natural scenes. In: Neural information processing systems, Vol 7. Morgan Kaufman, San Mateo

Bliss TVP, Collingridge GL (1993) A synaptic model of memory: long term potentiation in the hippocampus. Nature 361:31–39

Bliss TVP, Lomo T (1973) Long lasting potentiation of synaptic transmission in the dentate area of the anesthetized rabbit following stimulation of the perforant path. J Physiol [Lond] 232:331–356

Burgess N, Recce M, O'Keefe J (1995) Hippocampus – spatial models. In: The handbook of neural theory and neural networks, pp 468–472. Bradford Books/MIT Press, Cambridge, Mass

Buzsáki G (1989) A two-stage model of memory trace formation: a role for 'noisy' brain states. Neuroscience 31:551–570

Carpenter GA, Grossberg S (1993) Normal and amnesic learning, recognition and memory by a neural model of cortico-hippocampal interactions. Trends Neurosci 16:131–137

Christian EP, Dudek FE (1988) Electrophysiological evidence from glutamate microapplications for local excitatory circuits in the CA1 area of rat hippocampal slices. J Neurophysiol 59:110–123

Comon P (1994) Independent component analysis – a new concept? Signal Processing 36:287–314

Eichenbaum H, Otto T, Cohen NJ (1994) Two functional roles of the hippocampal memory system. Behav Brain Sci 17:449–518

Érdi P, Aradi I, Grőbler T (1997) Rhythmogenesis in single cells and population models: olfactory bulb and hippocampus. BioSystems 40:45–53

Fomin T, Kőrmendy-Rácz J, Lőrincz A (1997) Towards a unified model of cortical computation I. Data compression and data reconstruction using dynamic state feedback. Neural Network World 7:116–136

Freund TF, Buzsáki G (1996) Interneurons of the hippocampus. Hippocampus 6:347–470

Gluck MA (ed) (1996) Special issue. Computational models of hippocampal function in memory. Hippocampus 6:565–762

Harris EW, Cotman CW (1986) Long-term potentiation of guinea pig mossy fiber responses is not blocked by *N*-methyl D-aspartate antagonists. Neurosci Lett 70:132–137

Hwang YK, Ahuja N (1992) Gross motion planning – a survey. ACM Comput Surveys 24:219–291

Iijima T, Witter MP, Ichikawa M, Tominaga T, Kajiwara R, Matsumoto G (1996) Entorhinal-hippocampal interactions revealed by real-time imaging. Science 272:1176–1179

Jutten C, Herault J (1991) Blind separation of sources. part I. An adaptive algorithm based on neuromimetic architecture. Signal Processing 24:1–10

Kalmár Z, Szepesvári, C, Lőrincz A (1995) Generalized dynamic concept model as a route to construct adaptive autonomous agents. Neural Network World 5:353–360

Karhunen J, Oja E, Wang L, Vigário R, Joutsensalo J (1997) A class of neural networks for independent component analysis. IEEE Trans Neural Networks 8:486–504

Laheld B, Cardoso JF (1994) Adaptive source separation with uniform performance. In: Holt et al. (eds) Signal processing VII. Theories and applications. Proceedings of EUSIPCO-94, Edinburgh, UK, September 1994, Vol 2, pp 183–186

Li XG, Somogyi P, Ylinen A, Buzsáki G (1994) The hippocampal CA3 network: an in vivo intracellular labeling study. J Comp Neurol 339:181–208

Logothetis N, Pauls J, Poggio T (1995) Spatial reference frames for object recognition. Tuning for rotations in depth. A.I. Memo no. 1533

Lőrincz A (1997a) Hippocampal formation trains independent components via forcing input reconstruction. In: Gerstner W, Germond A, Hasler M, Nicoud J-D (eds) Artificial neural networks – ICANN 97, Lausanne, Switzerland. Springer, Berlin Heidelberg New York, pp 163–168

Lőrincz A (1997b) Towards a unified model of cortical computation. II. From control architecture to a model of consciousness. Neural Network World 7:137–152

Markram H, Tsodyks M (1997) The information content of action potential trains a synaptic basis. In: Gerstner W, Germond A, Hasler M, Nicoud J-D (eds) Artificial neural networks – ICANN 97, Lausanne, Switzerland. Springer, Berlin Heidelberg New York, pp 13–23

McClelland JL (1996) Role of the hippocampus in learning and memory: a computational analysis. In: Perception, memory and emotion: frontiers in neuroscience. Elsevier, Oxford, pp 601–613

O'Keefe J, Burgess N (1996) Spatial and temporal determinants of the hippocampal place cell activity. In: Perception, memory and emotion: frontiers in neuroscience. Elsevier, Oxford, pp 359–373

O'Keefe J, Nadel L (1978) The hippocampus as a cognitive map. Clarendon Press, Oxford

O'Keefe J, Recce ML (1993) Phase relationship between hippocampal place units and the EEG theta rhythm. Hippocampus 3:317–330

Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607–609

Ono T, McNaughton BL, Molotchnikoff S, Rolls ET, Nishijo H (1996) Perception, memory and emotion: frontiers in neuroscience. Elsevier, Oxford

Oram MW, Perrett DI (1996) Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. J Neurophysiol 76:109–129

Otto T, Eichenbaum H (1992) Complementary roles of orbital prefrontal cortex and the perirhinal-entorhinal cortices in an odor-guided delayed non-matching to sample task. Behav Neurosci 106:763–776

Rao RPN, Ballard DH (1997) Dynamic model of visual recognition predicts neural response properties in the visual cortex. Neural Comput 9:721–763

Rolls ET (1989) The representation and storage of information in neuronal networks in the primate cerebral cortex and the hippocampus. In: The computing neuron. Addison-Wesley, Wokingham, pp 125–129

Rolls ET (1996) The representation of space in the primate hippocampus, and episodic memory. In: Perception, memory and emotion: frontiers in neuroscience. Elsevier, Oxford, pp 375–400

Skaggs WE, McNaughton BL, Wilson MA, Barnes CA (1996) Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. Hippocampus 6:149–172

Squire LR (1992) Memory and the hippocampus: a synthesis of findings with rats, monkeys, and humans. Psychol. Rev. 99:195–231

Squire LR, Knowlton BJ (1995) Learning about categories in the absence of memory. Proc Natl Acad Sci USA 92:12470–12474

Szepesvári C, Lőrincz A (1996) Inverse dynamics controllers for robust control: consequences for neurocontrollers. In: Malsburg C von der, Seelen W von, Vorbrüggen JC, Sendhoff B (eds) Artificial neural networks – ICANN 96, Bochum, Germany. Springer, Berlin Heidelberg New York, pp 697–702

Szepesvári C, Lőrincz A (1997) Approximate inverse-dynamics based robust control using static and dynamic state feedback. In: Kalkkuhl J, Hunt KJ, Zbikowski R, Dzielinski A (eds) Applications of neural adaptive control technology. World Scientific, Singapore, pp 151–180

Szepesvári C, Cimmer S, Lőrincz A (1997) Neurocontroller using dynamic state feedback for compensatory control. Neural Networks 10:1691–1708

Tanaka K, Saito H, Funada Y, Moriya M (1991) Coding visual images of objects in the inferotemporal cortex of the macaque monkey. J Neurosci 6:134–144

Taylor JG (1991) Can neural networks ever be made to think? Neural Network World 1:4–11

Traub RD, Dingledine R (1990) Model of synchronized epileptiform bursts induced by high potassium in CA3 region of rat hippocampal slice: Role of spontaneous EPSPs in initiation. J Neurophysiol 64:1009–1018

Tsodyks MV, Skaggs WE, Sejnowski TJ, McNaughton BL (1996) Population dynamics and theta rhythm phase precession of hippocampal place cell firing: a spiking neuron model. Hippocampus 6:271–280

Wang L, Karhunen J, Oja E (1995) A bigradient optimization approach for robust PCA, MCA, and source separation. In:

Proc. IEEE Int. Conf. Neural Networks, Perth, Australia, November 1995. IEEE Publishing, pp 1684–1689

Wittmeyer H (1936) Ueber die Loesung von linearen Gleichungssystemen durch Iteration. Z Angew Math Mech 16:301–310

Young B, Eichenbaum H (1996) What do hippocampal neurons encode? In: Perception, memory and emotion: frontiers in neuroscience. Elsevier, Oxford, pp 229–249

Zola-Morgan S, Squire LR (1991) The primate hippocampal formation: evidence for a time-limited role in memory storage. Science 250:288–289

Zola-Morgan S, Squire LR, Rempel NL, Clower RP, Amaral DG (1992) Enduring memory impairment in monkeys after ischemic damage to the hippocampus. J Neurosci 12:2582–2596