Attentional filtering in neocortical areas: A top-down model

András Lőrincz

Eötvös Loránd University, 1117 Budapest, Pázmány Péter sétány 1/D, Hungary

Abstract

Two comparator based rate code models – a control architecture and the corresponding reconstruction network – are merged. The role of reconstruction is information maximization and bottom-up (BU) noise filtering, whereas top-down control 'paves the way' for context compliant BU information that, together, we identify with attentional filtering. Falsifying prediction of the model has gained experimental support.

Key words: attention, neocortex, top-down control, reconstruction network

1 Introduction

We start from a comparator-based control model [1], review the related reconstruction network model [2], map the combined network onto the neocortex, and model attentional filtering.

2 Model description

The controller. Our control model is formulated in terms of speed-field, that is, state dependent directions pointing towards target positions. A particular speed-field is given, for example, by the difference vectors between the target state and all other states. The control task, called speed-field tracking, is defined as moving according to the speed-field at each state (see, e.g., [1] and references therein for the mathematical details). The dynamic equation of a system is a set of continuous differential equations that determines the change of state per unit time $\dot{\mathbf{x}} \approx \frac{\Delta \mathbf{x}}{\Delta t}$ given the state of the 'plant' $\mathbf{x} \in \mathrm{R}^n$ and the external forces acting upon that plant, including the control action $\mathbf{u} \in \mathrm{R}^p$. Let $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u})$ denote the dynamics, i.e., what the plant does.

Inverse dynamics works in the opposite way: given the state and the desired change of state, inverse dynamics provides the control vector. Let $\mathbf{v}(\mathbf{x}) \in \mathrm{R}^n$ denote the *desired* speed. Assume that we have an approximate model of the precise inverse dynamics $\mathbf{u}_{ff} : \hat{\mathbf{u}}_{ff}(\mathbf{x}, \mathbf{v}(\mathbf{x})) = \hat{\mathbf{A}}(\mathbf{x})\mathbf{v}(\mathbf{x}) + \hat{\mathbf{b}}(\mathbf{x})$. If $\hat{\mathbf{u}}_{ff}$ influences the plant directly, it is called *feedforward controlling*. For perfect inverse dynamics, the control vector produces the desired speed: $\mathbf{v}(\mathbf{x}) = \mathbf{f}(\mathbf{x}, \mathbf{u}_{ff}(\mathbf{x}, \mathbf{v}(\mathbf{x})))$. If the feedforward control vector is imprecise,

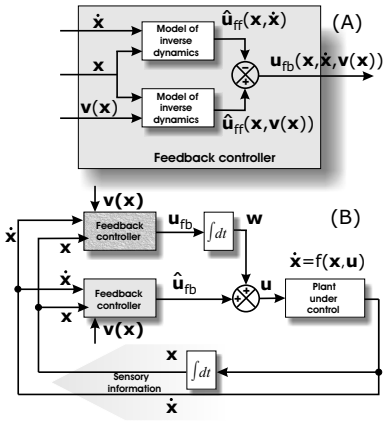

Fig. 1. **Robust controller for speed-field tracking tasks**
A: Model of the inverse dynamics. **B:** SDS feedback controller. (See text.)

then comparison error $\mathbf{e}_c = \mathbf{v}(\mathbf{x}) - \dot{\mathbf{x}}$, the difference between desired and experienced speeds appears. A model of the inverse dynamics can be used to correct this error. The error correcting controller, called *feedback controller*, accumulates error correcting control actions. Our model assumes the form $\mathbf{u}_{fb} = \hat{\mathbf{B}}(\mathbf{x})(\mathbf{v}(\mathbf{x}) - \dot{\mathbf{x}})$. The control vector is built from these two terms $\mathbf{u}(t) = \hat{\mathbf{u}}_{ff}(t) + \mathbf{w}(t) = \hat{\mathbf{u}}_{ff}(t) + \int_{-\infty}^{t} \mathbf{u}_{fb}(t')\, dt'$. Our *key observation* (Fig. 1) is that the feedforward controller can be a comparator, too [1]: $\hat{\mathbf{u}}_{ff} = \hat{\mathbf{u}}_{fb} = \hat{\mathbf{A}}(\mathbf{x})(\mathbf{v}(\mathbf{x}) - \dot{\mathbf{x}})$. Taken together:

$$\mathbf{u} = \hat{\mathbf{u}}_{fb} + \int_{-\infty}^{t} \mathbf{u}_{fb}(t')\, dt' = \hat{\mathbf{A}}(\mathbf{x})(\mathbf{v}(\mathbf{x}) - \dot{\mathbf{x}}) + \int_{-\infty}^{t} \hat{\mathbf{B}}(\mathbf{x})(\mathbf{v}(\mathbf{x}) - \dot{\mathbf{x}}). \qquad (1)$$

Both controllers compare 'desired' and 'experienced' quantities, which makes a useful compromise. Control may not be perfect, but explicit modelling of the highly non-linear term $\mathbf{b}(\mathbf{x})$ is not necessary, because it does not appear in the differences. The architecture, the *static and dynamic state* (SDS) feedback controller, is globally stable under mild conditions: $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ needs to be 'sign-proper': only the proper sign of each control component needs to be warranted.

The reconstruction network. The basic reconstruction network (Fig. 2A) can be built the same way if the input and the reconstructed input of the reconstruction process are identified with the 'desired' and 'experienced' quantities, respectively. The minimal network has two layers: the reconstruction

error layer computes the difference ($\mathbf{e} \in \mathrm{R}^r$) between input ($\mathbf{x} \in \mathrm{R}^r$) and reconstructed input ($\mathbf{y} \in \mathrm{R}^r$): $\mathbf{e} = \mathbf{x} - \mathbf{y}$. The reconstructed input $\mathbf{y}$ is produced by top-down (TD) transformation matrix $\mathbf{Q} \in \mathrm{R}^{r \times s}$ inputted by the hidden representation $\mathbf{h} \in \mathrm{R}^s$. The hidden representation is corrected by BU transformed reconstruction error, i.e., by $\mathbf{We}$ ($\mathbf{W} \in \mathrm{R}^{s \times r}$ of rank $\min(s, r)$). Correction means that the value of the hidden representation is maintained and the correcting amount is added. In turn, the hidden representation needs self-excitatory connections ($\mathbf{M}$) to sustain the activities. For input $\mathbf{x}$, the iteration stops when $\mathbf{WQh} = \mathbf{Wx}$. That is, to each input, the relaxed hidden representation is solely determined by TD matrix $\mathbf{Q}$, which is identified with the long-term memory. BU matrix $\mathbf{W}$ is perfectly tuned if $\mathbf{W} = (\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T$, i.e., if $\mathbf{WQ} = \mathbf{I}$ ($\mathbf{I} \in \mathrm{R}^{s \times s}$). Then the network is as fast as feedforward nets.

For reconstruction, the network should optimize BU information transfer (Fig. 2(B)): an additional layer is introduced to hold independent components (IC) $\mathbf{s} = \mathbf{We}$ to meet this constraint. ICs enable local noise filtering if reconstructed input and error layers are separated and there are two separate afferents to the IC layer: one carrying the error, and another ($\mathbf{P}$) carrying the reconstructed input. Nonlinearity in the latter can locally remove input noise. It is alike to wavelet denoising, but instead of wavelet transformation, BU transformation is optimized for the inputs experienced by the network. The method is called sparse code shrinkage (SCS) [3]. SCS concerns the components of the BU transformed reconstructed input: high (low) amplitude components of the BU transformed reconstructed input $\mathbf{Py}$ can (not) open the gates of components of the IC layer and IC transformed reconstruction error can (not) pass the open (closed) gates to correct the hidden representation. Note that $\mathbf{P} = \mathbf{W}$ for well tuned system [4]. Also, the transformation $\mathbf{s} \rightarrow \mathbf{h}$, i.e., $\mathbf{N}$ can support pattern completion [2]. Apart from SCS, the reconstruction network is linear, which property is denoted by sign '$\sim$'. '$A \sim B$' means that up to a scaling matrix, quantity A is approximately equal to quantity B. For a *well tuned network* and if matrix $\mathbf{M}$ performs temporal integration, $\dot{\mathbf{x}} \cong \mathbf{e} \sim \mathbf{s}$ by construction. Similarly, hidden representation $\mathbf{h} \sim \mathbf{y}$, apart from noise, $\mathbf{y} \sim \mathbf{x}$ and $\mathbf{y} = \int \mathbf{e}_f dt$, $\mathbf{e}_f$ is the noise filtered reconstruction error $\mathbf{e}$.

The joined model. The reconstruction network can be controlled (Fig. 2B,C). Control influences the hidden layer by $\sim \dot{\mathbf{v}}$. Temporal integration occurs at the hidden layer and the effect of the controller could be equal to $\mathbf{Av}$. However, the correct form of matrix $\mathbf{A}$ is not known and the SDS controller is used to achieve approximately perfect control. The controller is made of another reconstruction network $\mathcal{H}$, the 'higher' network, which receives input from the IC layer of the 'lower' controlled network $\mathcal{L}$ (Fig. 2D). Control works by subtracting the desired speed from the input of the higher network. The input to network $\mathcal{H}$ is equal to $\dot{\mathbf{x}} - \mathbf{v}(\mathbf{x})$ (via $\mathbf{C}_1$ and from outside, respectively) that can be modulated by $\mathbf{x}$ via $\mathbf{C}_2$. By construction, (i) the input is noise filtered and reconstructed, the reconstructed input is $\sim (\dot{\mathbf{x}} - \mathbf{v}(\mathbf{x}))$ and (ii) apart from

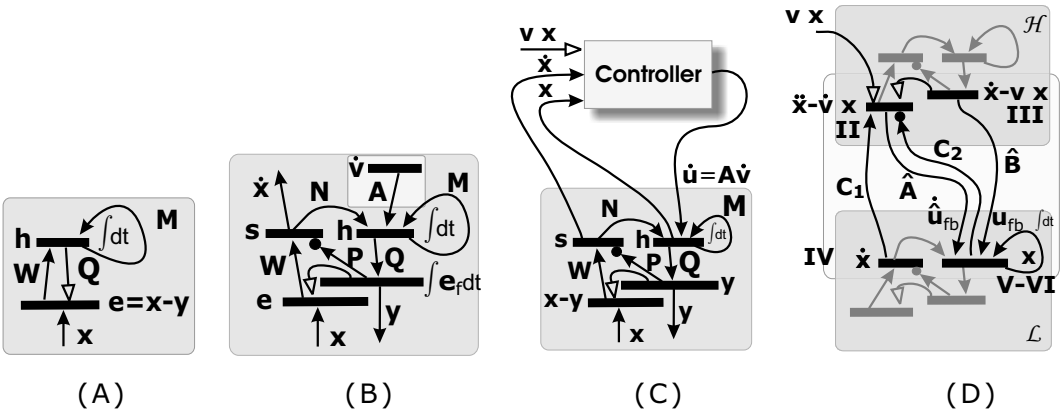

Fig. 2. **Reconstruction networks (A and B) with controllers (C and D)** **A:** Simple reconstruction network (RCN). **B:** RCN with local noise filtering and under top-down influence. **C:** Controlled RCN. **D:** Controller is built from another RCN. Roman numbers: mapping to neocortical layers. Arrows with dots: nonlinear modulation. (See text.)

the noise content, the error vector approximates the temporal derivative of the reconstructed input. These two differences undergo linear transformations and are added to the hidden representation of network $\mathcal{L}$, where – by construction – they undergo temporal integration. The resulting signals are the SDS control signals, which enforce the desired speed onto lower network $\mathcal{L}$ under the mild condition that control is 'sign proper'.

3 Discussion

The reconstruction network (Fig. 2A) was suggested by Kawato et al., to model corticocortical connections [5]. Relations to other models can be found therein. A more sophisticated version was reached by extending the description of the hippocampal-entorhinal loop [4] to model the neocortex [2]. Here, intralaminar connections and attention are modelled by using control principles [1].

The perfectly tuned architecture behaves as a BU feedforward network and can be biased by TD influence. Consider lower reconstruction network $\mathcal{L}$ of Fig. 2(D). The bias will modify the hidden representation of the network $\mathcal{L}$, which may or may not fit the input. If it does not fit, then reconstruction error appears but only a small portion of this error can pass the sparsification process at the IC layer, because of SCS thresholding. That is, information that matches the *context* of the higher reconstruction network will be able to pass sparsification, whereas all other information will be attenuated by the SCS. In turn, TD influence 'paves the way' of some of the components. This process can be seen as *attentional filtering*. For supporting computational demonstrations, see [6].

Now, we shall map the architecture onto the six-layered neocortex. Most prominent neocortical connections are depicted in Fig. 2(D) within the light gray box between networks $\mathcal{H}$ and $\mathcal{L}$. Roman numbers denote neocortical layers. Input arrives at layer IV ($\sim \dot{\mathbf{x}}$). Layer IV neurons send messages to layer II ($\ddot{\mathbf{x}} - \dot{\mathbf{v}}(\mathbf{x})$) and layer III ($\dot{\mathbf{x}} - \mathbf{v}(\mathbf{x})$) and also to layer VI. Superficial neurons provide output down to layers V-VI ($\sim \mathbf{x}$). Layers V-VI provide feedback to layers II and III. The main output to higher cortical areas emerges from layers II-III. The main feedback to lower areas is provided by layer V [7]. The theoretical model and the anatomical structure can be matched by assuming that superficial layers of the lower cortical layer *and* deep layers of the higher cortical layer form *one* functional unit, the reconstruction network [2], whereas the interlaminar connections correspond to the control architecture.

A falsifying prediction of the model concerns the hidden representation, which has to maintain its own activities in order to enable additive corrections and temporal integration. Persistent activities in the deep layers but not in the superficial layers of the entorhinal cortex have been found experimentally [8], providing support to our model. Another falsifying prediction of the joined model is that TD connections between neocortical layers can be interpreted as long-term memories (see also [2]), because these connections are responsible for the relaxed activities of the hidden layers. These connections are generally more numerous than the feedforward connections between the same areas, but the activity flow along these connections is relatively low and suggests a weak functional role [7]. This apparent discrepancy may be resolved by noting that different interpretations may coexist in the brain as it has been made evident, e.g., in the animal experiments on binocular rivalry [9]. If reconstruction concerns a single interpretation, then feedback activity flow should be *small*: There are evidences that activities in V4 (responsible for conscious detection of colors) and V5 (responsible for conscious detection of fast motion) in the monkey are uncorrelated. According to Zeki [10], uncorrelated activities indicate that conscious experiences propagate downwards along parallel channels. Moreover, the conscious binding of the result of the individual conscious experiences seems to be delayed [11]. In turn, it is possible that only a single interpretation – and a sparse activity pattern – is communicated downwards at a time.

According to recent measurements, awareness and attention needs to be distinguished [12] and attention increases neuronal activities responsible for the processing of the attended stimuli. Most probably, endogenous attention facilitates the pathways that should be used by the attended stimuli [13]. In our model, facilitation can manifest itself through control action *within* cortical layers. On the other hand, it has been argued that awareness involves recurrent interactions between areas and may be suppressed by backward masking [12]. This recurrent interaction required for awareness is our candidate function of the feedback corticocortical connections.

Conclusions. A model of neocortical information processing has been presented and mapped onto the neocortex. The model suggests that bottom-up noise filtering is accomplished by reconstruction networks *between* neocortical layers, whereas top-down modulation of this filtering is the task of the cortical layers. The model provides explanation on feedback connections between cortical layers, which are more numerous than the bottom-up connections between the different areas, but are very quiet: These connections are the long-term memories of the model, which communicate a single interpretation at a time. The model allowed us to distinguish between attention and awareness, two delicate and intertwined concepts.

References

[1] A. Lőrincz, Static and dynamic state feedback control model of basal ganglia - thalamocortical loops, Int. J. of Neural Systems 8 (1997) 339–357.

[2] A. Lőrincz, B. Szatmáry, G. Szirtes, Mystery of structure and function of sensory processing areas of the neocortex: A resolution, J. Comp. Neurosci. 13 (2002) 187–205.

[3] A. Hyvärinen, Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation, Neural Computation 11 (1999) 1739–1768.

[4] A. Lőrincz, G. Buzsáki, Two–phase computational model training long–term memories in the entorhinal–hippocampal region, NYAS 911 (2000) 83–111.

[5] M. Kawato, H. Hayakawa, T. Inui, A forward-inverse model of reciprocal connections between visual neocortical areas, Network 4 (1993) 415–422.

[6] B. Takács, A. Lőrincz, Supporting material on neocortical attentional control, TechRep: http://people.inf.elte.hu /lorincz/Files/CNS04/CNS04Support.pdf.

[7] E. M. Callaway, MIT Encycl. of Cogn. Sci., MIT Press, Cambridge, MA, 2000, Ch. Visual cortex, cell types and connections in, pp. 867–869.

[8] A. V. Egorov, B. N. Hamam, E. Fransén, M. E. Hasselmo, A. A. Alonso, Graded persistent activity in entorhinal cortex neurons, Nature 420 (2002) 173–178.

[9] D. A. Leopold, N. K. Logothetis, Multistable phenomena: Changing views in perception, Trends in Cognitive Sciences 3 (1999) 254–264.

[10] S. Zeki, The disunity of consciousness, Trends in Cogn. Sci. 7 (2003) 214–218.

[11] A. Bartels, S. Zeki, The temporal limits of binding: Is binding post-conscious?, in: Soc. Neurosci. Abstr., Vol. 11, 2002, p. 260.

[12] V. Lamme, Why visual attention and awareness are different?, Trends in Cogn. Sci. 7 (2003) 12–18.

[13] H. Egeth, S. Yantis, Visual attention: Control, representation, and time course, Annual Review of Psychology 48 (1997) 269–297.