Object Localisation Using Laterally Connected "What" and "Where" Associator Networks

C. Weber and S. Wermter

Hybrid Intelligent Systems Group, School of Computing and Technology, University of Sunderland, Sunderland SR6 0DD, United Kingdom [www.his.sunderland.ac.uk]

E-mail: {Cornelius.Weber, Stefan.Wermter}@Sunderland.ac.uk

Abstract. We describe an associator neural network to localise a recognised object within the visual field. The idea extends the use of lateral connections within a single cortical area to their use between different areas. Previously, intra-area lateral connections have been implemented within V1 to endow the simple cells with biologically realistic orientation tuning curves as well as to generate complex cell properties. In this paper we extend the lateral connections to also span an area laterally connected to the simulated V1. Their training was done by the following procedure: every image on the input contained an artificially generated orange fruit at a particular location. This location was reflected – in a supervised manner – as a Gaussian on the area laterally connected to V1. Thus, the lateral weights are trained to associate the V1 representation of the image to the location of the orange. After training, we present an image with an orange of which we do not know its location. By the means of pattern completion a Gaussian hill of activation emerges on the correct location of the laterally connected area. Tests display a good performance with real oranges under diverse lighting and backgrounds. A further extension to include multi-modal input is discussed.

Introduction

Once that an object of interest appears in the visual field, it is necessary to localise its position within the visual field before moving the centre of sight toward it and, eventually, to activate a grasping movement prototype [9]. We develop a biologically inspired solution using a recurrent associator network which we want to apply in a bio-mimetic mirror neuron-based robot, MirrorBot.

Our approach extends the framework of intrinsic lateral (horizontal) connections in the cortex toward object recognition and localisation. Horizontal connections within one cortical area have a strong influence on cortical cell response properties. In the visual area V1, for example, they may be responsible for surround effects and for the non-linear response properties of simple and complex cells [11]. This view is supported by connectionist neuron learning paradigms in which lateral connections statistically de-correlate [10] or find correlation structure [1] within the activities of cells in an area. Both paradigms are in accordance with the notion that the lateral connections form an attractor network. The activation patterns which form its attractors correlate nearby cell's activations

but de-correlate distant cell's activations. The attractor activation pattern can recover noisy input with maximum likelihood [2]. Such a theoretically derived learning paradigm has successfully explained orientation tuning curves of V1 simple cells as well as complex cell's response properties [13].

Here we apply the learning rule for lateral connections within a cortical area to connections between different, but laterally organised cortical areas. This is justified by the fact that lateral connections between areas – as opposed to hierarchical connections – originate and terminate in the same cortical layers [3]. A different learning rule is applied to the hierarchical connections which form the input to one of our two simulated laterally connected areas (see Fig. 1). This is a rule which leads to feature extraction and can be any rule from the sparse coding / ICA repository. Here we use a sparse coding Helmholtz machine for the bottom-up connections, as previously described [13].

The two laterally connected areas of our model specialise on object recognition and localisation. As such they shall be regarded as exemplary areas within the lateral "what" and the dorsal "where" pathway of the visual system. In the actual implementation, however, in a model where every connection is trained and which uses natural images as input, there are no high-level cortical areas. Instead, our "what" area receives direct visual input, reminiscent of V1 while our "where" area receives directly the representation of a location. Such a representation may actually reside in the superior colliculus [5].

The problem of object localisation is intermixed with recognition: several structures in different locations within the image may match to the object of interest and the best matching location has to be found. For this purpose, saliency maps can be produced [8] or the data may be generated from Bayesian priors [12]. These approaches, however, are missing a neural description. An approach involving shifter neurons [7] takes into consideration the derivative of an object with respect to a shift within the image. It can handle small shifts of complex objects but involves high dimensional neurons which each have an $N \times N$ matrix to the N input neurons. Our approach uses standard neurons with order N connections and handles relatively large shifts. However, tests have been done only with a very simple object, and an extension to general objects is discussed.

Theory and Methods

The architecture is depicted in Fig. 1 and consists of a "what" pathway on the left, and a "where" pathway on the right. The "what" pathway consists of an input area and a hidden area. The input area consists of three sub-layers to receive the red, green and blue components of colour images. Its size of 24×16 pixels which is minimal to demonstrate object localisation reflects the computational demand of training. The hidden area of the "what" pathway consists of two layers which we may loosely identify with the simple (lower layer) and complex (upper layer) cells of V1. The lower layer receives bottom-up connections W^{bu} from the input. In the following we will assume that these have already been trained such that the lower layer cells resemble the simple cells of V1 [13]. Since

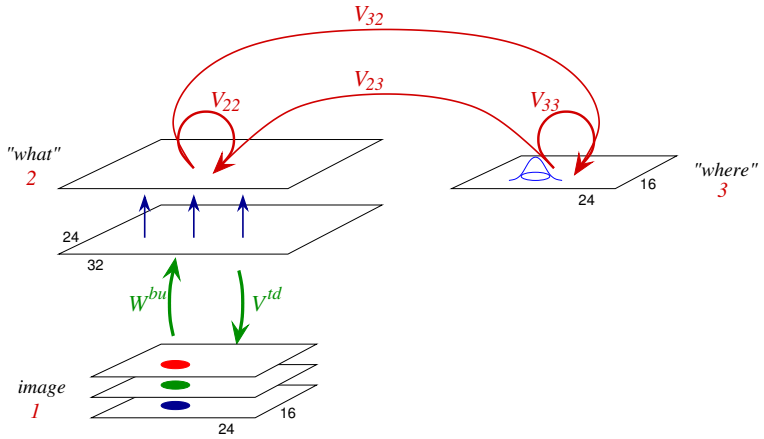

Fig. 1. Model architecture. Left, the pathway of the lower visual system, the retina which receives the image and V1, which we refer to as the "what" area. Feature extracting, hierarchically organised weights are W^{bu}, V^{td} (green). On the right side, the "where" area displays the location of the object of interest. Lateral association weights are V_{22}, V_{33}, V_{23} and V_{32} (red). The small numbers denote simulated area sizes.

colour images are used, a few cells have learnt to encode colour, while the majority has become black-and-white edge detectors. The depicted top-down weights V^{td} were only used to train W^{bu}, but are not used further on. The upper layer of the V1 cells receives a copy of the output of the lower layer cells. After it receives this initial input, it functions as an attractor network which solely updates its activations based on its previous activations. Each cell receives its input from all other neurons via recurrent weights V_{22}. In addition, input arrives from the laterally connected area of the "where" pathway via weights V_{23}.

The "where" pathway on the right of Fig. 1 consists of just one area. Its size of 24×16 neurons matches the size of the image input area of the "what" path, because an interesting object within the image should have a representation as an activation at the corresponding location on the "where" area. The "where" neurons are fully connected via recurrent weights V_{33} and in addition receive input from the highest "what" layer via V_{32}. In the following, we will refer to all connections V_{22}, V_{33}, V_{23} and V_{32} collectively as V^{lat}, because they always receive the same treatment, during training as well as during activation update.

Activation Dynamics and Learning rule: The activation update of the "where" and highest level "what" neurons is governed by the following equation:

$$u_i(t+1) \;=\; f(\textstyle\sum_l v_{il}^{lat} u_l(t)) \qquad (1)$$

Activation u_i of neuron i develops through discrete time t using the input via lateral weights v_{il}^{lat} from the other l ("what" and "where") neurons. The lateral weights are not forced to be symmetric, i.e. $v_{il}^{lat} \neq v_{li}^{lat}$ in general.

The lateral weights are trained from the bottom-up input. Their purpose is to memorise the incoming activities $u_i(t = 0)$ as activation patterns which

they maintain. Since they will not be capable of holding every pattern, they will rather classify these into discrete attractors. In the original top-down generative model [13] these patterns were recalled in a separate mode of operation ("sleep phase") in order to generate statistically correct input data.

Learning maximises the log-likelihood to generate the incoming data distribution by the internal activations $u_i(t)$ if Eq. 1 is applied repeatedly:

$$\Delta v_{il}^{lat} \approx \sum_t (u_i(t=0) - u_i(t))\, u_l(t-1)\,. \tag{2}$$

Transfer Function and Parameters: The transfer function of our continuous rate-coding neurons is:

$$f(h_i) = \frac{e^{\beta h_i}}{e^{\beta h_i}+n} \approx p_i(1) \tag{3}$$

The function ranges from 0 to 1 and can be interpreted as the probability $p_i(1)$ of a binary stochastic neuron i to be in active state 1. Parameters $\beta = 2$ scales the slope of the function and n is the degeneracy of the 0-state. Large $n = 8$ reduces the probability of the 1-state and accounts for a sparse representation of the patterns which are learned. The introduction of this sparse coding scheme was found to be more robust than the alternative use of variable thresholds. The weights V^{lat} were initialized randomly, self-connections were constrained to $v_{ii}^{lat} = 0$.

Training Procedure: First, the weight matrices W^{bu} and V^{td} were trained on small patches randomly cut out from 14 natural images, as in [13], but with a 3-fold enlarged input to separate the red, green and blue components of each image patch. 200000 training steps had been done. Lateral weights V^{lat} were then trained in another 200000 training steps with W^{bu} and V^{td} fixed. Herefore, within each data point (an image patch), an artificially generated orange fruit was placed to a randomly chosen position. An orange consisted of a disc of 5 pixels in diameter which had a color randomly chosen from a range of orange fruit photos. The mean of the pixel values was subtracted and the values normalised to variance 1. The "where" area received a Gaussian hill of activity on the location which corresponds to the one in the input where the orange is presented. The standard deviation of the Gaussian hill was $\sigma = 1.5$ pixels, the height was 1.

The representation of the image with an orange obtained through W^{bu} on the lower V1 cells was copied to the upper V1 cells. This together with the Gaussian hill on the "where" area was used as initial activation $u_i(t = 0)$ to start the relaxation procedure described in Eq. 1. It is also used as target training value. Relaxtions were done for $0 \leq t < 4$ time steps.

Results

Anatomy: Fig. 2 a) shows a sample of weights W^{bu} of our lower V1 cells. Many have developed localized, Gabor function shaped, non color selective receptive fields to the input. A few neurons have developed broader, color selective receptive fields. Similar results have been obtained [4].

Fig. 2 b)-e) shows samples of the lateral connections V^{lat}. Inner-area connections are usually center-excitatory and surround inhibitory in the space of their

a) from W_{10} **b)** from V_{22} **c)** from V_{23} **d)** from V_{32} **e)** from V_{33}

Fig. 2. a) The receptive fields (rows of W^{bu}) of 16 adjacent lower V1 ("simple") cells. Bright are positive, dark negative connection strengths to the red, green and blue visual input. Receptive fields of color selective neurons appear colored, because the three color components differ. **b)-e)** Samples of lateral weights V^{lat}. Positive connections are green, negative are red. **b)** Within-area lateral connections among the upper V1 ("complex") cells. **c)** Lateral cross-area connections from the "where" area to upper V1 to the same 16 neurons (same indices) as depicted in a) and b). Connections V_{22} and V_{23} together form the total input to an upper V1 cell. **d)** Cross-area lateral connections from upper V1 to the "where" area. **e)** Within-area lateral connections on the "where" area to the same 16 neurons as depicted in d). Connections V_{33} and V_{32} together form the total input to a "where"-area cell. Within-area connections are in general center-excitatory and surround-inhibitory and they are small in the long range. Connections V_{33} establish a Gaussian-shaped hill of activations. Cross-area connections V_{32} influence the position of the activation hill. Self-connections in V_{22} and V_{33} are set to zero.

functional features [13]. Cross-area connections are sparse and less topographic. Strong connections are between the "where" cells and color selective "what" cells, because for orange fruits, color is a salient identification feature.

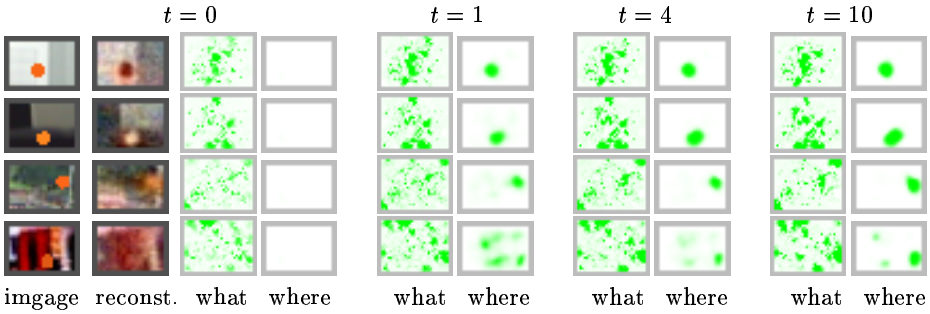

Fig. 3. Each row shows the network response to a color image which contains an artificially generated orange fruit. From left to right: the image, the reconstruction of the image using feedback weights V^{td}, the representation on the "what" area, the initial zero activities on the "where" area at time $t = 0$. Then the activations on the "what" and "where" areas at time $t = 1$, then on both at time $t = 4$ which is the relaxation time used for training, and then after a longer relaxation of $t = 10$ time steps.
The estimated position of the orange on the "where" area is correct in the upper 3 rows. In the difficult example below, at time $t = 1$ activity on the "where" area is distributed across many locations and later focuses on a wrong location.

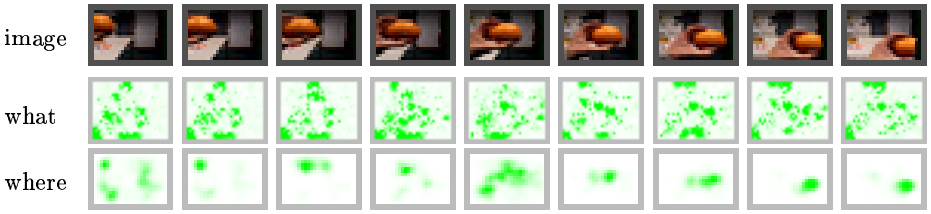

Fig. 4. Localisation on real images taken from the robot camera. The lower two rows show the response on the "what" and the "where" area at iteration time $t = 4$ to the image in the upper row.

Physiology: Figs. 3 and 4 show the relaxation of the network activities after initialization with sample stimuli. In all cases, the "where" area neuron's activations were initialised to zero at time $t = 0$. The relaxation procedure therefore completes a pattern which spans both, the "what" and the "where" area, but which is incomplete at time $t = 0$, as can be seen in Fig. 3.

The activition on the "where" area may resemble a Gaussian already at time $t = 1$, even though at this time, no effective input from the lateral weights V_{33} has arrived. A clearer Gaussian hill of activity evolves at later steps, but since no new information is coming in, the competition may draw a wrong location as a winner, if the representation is very fuzzy initially (Fig. 3, lowest row). Since the attractor shares the "what" and the "where" area, the Gaussian on the "where" area may remain distorted for quite a while.

All weights in the model have been trained on the basis of real images and are therefore irregular. Localisation quality may vary at slightly different object locations within the image. The 5th frame in Fig. 4, for example, leads to an unclean "where" representation. If information from the 4th frame would be taken into account, this may be cleand up. However, for simplicity and consistency with the training procedure, the algorithm processes only one frame at a time.

Fig. 5 shows how the network creates images, if there is no information but the location of the orange fruit. The projection of the corresponding internal representation onto the input creates images with predominantly blue background and a large patch of orange/red color near the location of the imaginated orange.

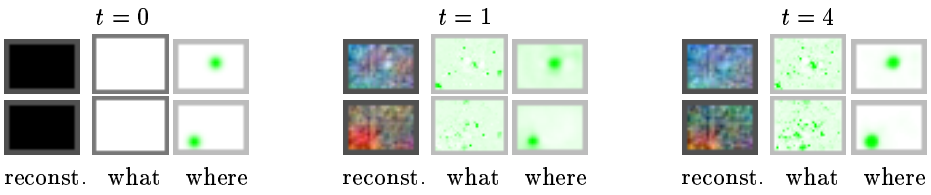

Fig. 5. Each row shows the network response purely from a Gaussian hill of activation on the "where" area. At time $t = 0$ the "what" area does not contain any activations and the reconstructed image is empty. Later, the areas maintain stable activations.

Discussion and Future Work

The current model has been trained on one object type, orange fruits. The cross-area lateral connections V_{32} originate predominantly at color selective V1 neurons, taking advantage of a feature specific to our chosen kind of object. For general object localisation, the cross-area lateral connections V_{32} need to be unspecific to object features. Then the object to localise would have to show up on V1 as a region of increased activation (attention). Fig. 6 shows two conceptual architectures which could achieve this. In both cases a third area, e.g. a language area, connects to V1. If it currently represents a specific object (as an orange in the figure) then it shall give an activation bias to those neurons on V1 which represent that object. Then, the lateral connections V_{32} transfer the biased representation to the "where" area, where intra-area connections V_{33} confine the activations to a Gaussian hill on the corresponding position. Note that direct connections between the language area and the "where" area would not make any sense, because an object does not have a preferred position *a priori*. Training would automatically lead to near-zero connections.

The question remains whether the language area should be connected laterally to the V1 area as in Fig. 6 **a)** or hierarchically as in Fig. 6 **b)**. The weight structure is the same, but their usage and training differs. In the lateral version a full representation on all three upper areas has to be present during training. In the hierarchical version the orange/apple features on the now highest level may be extracted by unsupervised training. The latter version is more appealing, because we expect a more abstract object representation to be hierarchically higher than a representation which still contains information on the object location.

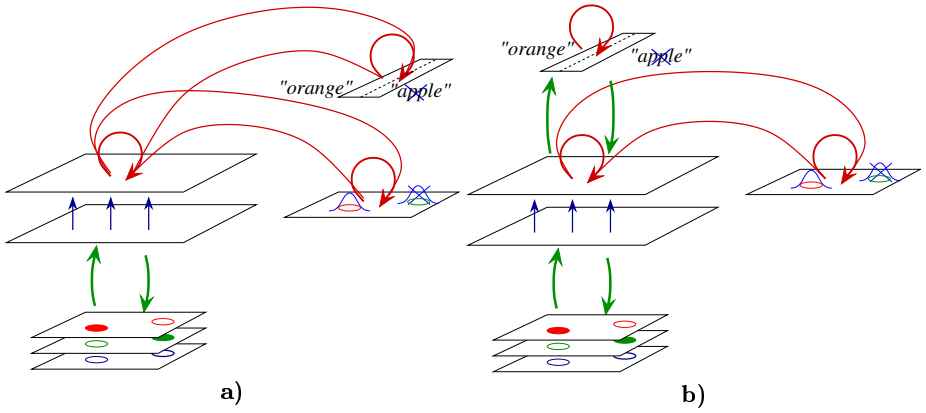

Fig. 6. The model with **a)** a laterally and **b)** a vertically (hierarchically) connected language area. If two trained fruit, like a red orange and a green apple, appear in the visual input, below, then at both corresponding locations in the "where" area a Gaussian may emerge, both competing. Input from another area, like the word "orange" from a language area may then strengthen the corresponding representation on "V1".

Both paradigms, however, are not necessarily contradicting: in the visual system, hierarchical and lateral connection patterns coexist if the vertical hierarchical level difference is small [3]. Two hierarchically arranged areas (with asymmetric connectivity [3]) may therefore use their mutual lateral connections (which are symmetric) for a "top-down" reconstruction. In addition, recent evidence suggests that object representations are distributed across different areas [6], potentially on different hierarchical levels. A model for a data driven arrangement of areas in parallel or hierarchically has been presented [14].

Acknowledgments

This work is part of MirrorBot, a project supported by the EU in the FET-IST programme under grant IST-2001-35282, coordinated by Prof. Wermter.

References

1. P. Dayan. Recurrent sampling models for the Helmholtz machine. *Neur. Comp.*, 11:653–77, 2000.
2. S. Deneve, P.E. Latham, and A. Pouget. Reading population codes: a neural implementation of ideal observers. *Nature Neurosci.*, 2(8):740–5, 1999.
3. D.J. Felleman and D.C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.
4. P.O. Hoyer and A. Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Comput. Neural Syst.*, 11(3):191–210, 2000.
5. A.K. Moschovakis, G.G. Gregoriou, and H.E. Savaki. Functional imaging of the primate superior colliculus during saccades to visual targets. *Nature Neurosci.*, 4(10):1026–31, 2001.
6. F Pulvermüller. A brain perspective on language mechanisms: from discrete neuronal ensembles to serial order. *Progress in Neurobiology*, 67:85–111, 2002.
7. R.P.N. Rao and D.H. Ballard. Development of localized oriented receptive fields by learning a translation-invariant code for natural images. *Network: Comput. Neural Syst.*, 9:219–37, 1998.
8. R.P.N. Rao, G.J. Zelinsky, M.M. Mayhoe, and D.H. Ballard. Eye movements in iconic visual search. *Vis. Res.*, 42(11):1447–63, 2002.
9. G. Rizzolatti and G. Luppino. Cortical motor system. *Neuron*, 31:889–901, 2001.
10. J. Sirosh and R. Miikkulainen. Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. *Neur. Comp.*, 9:577–94, 1997.
11. J. Sirosh, R. Miikkulainen, and Y. Choe, editors. *Lateral Interactions in the Cortex: Structure and Function.* Hypertext Book, www.cs.utexas.edu/users/nn/webpubs/htmlbook96, 1996.
12. J. Sullivan, A. Blake, M. Isard, and J. MacCormick. Bayesian object localisation in images. *Int. J. Computer Vision*, 44(2):111–35, 2001.
13. C. Weber. Self-organization of orientation maps, lateral connections, and dynamic receptive fields in the primary visual cortex. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Proceedings ICANN*, pages 1147–52, 2001.
14. C. Weber and K. Obermayer. *Emergent Neural Computational Architectures*, chapter Emergence of modularity within one sheet of neurons: a model comparison, pages 53–76. Springer-Verlag Berlin Heidelberg, 2001.