Development of Representations, Categories and Concepts—a Hypothesis

Harri Valpola
Laboratory of Computational Engineering
Helsinki University of Technology
FI-02150 TKK, Finland
Harri.Valpola@tkk.fi

***Abstract*— A long-standing question in cognitive sciences and machine learning is how a system can develop high-level concepts and categories which are useful for motor and cognitive control. I propose an architecture which learns a hierarchy of increasingly abstract, invariant features. Invariance is achieved by selecting information which reflects distinctions present in supervisory signals conveyd by contextual inputs. The main hypothesis is that the right contextual information can be efficiently distributed by associations and attentional process. The original sources of contextual information are specialised systems which reflect the innate, hard-wired behavioural goals of the system. Sensorimotor coordination generates structured sensory stimuli and the intrinsic contextual signals can select the behaviourally significant structures.**

***Index Terms*— Learning, representations, behaviour, invariance, perception**

I. Introduction

Complex behaviour requires efficient representations of the goals, actions and environment of the behaving system. A key issue is how useful representations, concepts and categories—such as “food”, “chair” and “heavy”—are acquired.

Unsupervised learning is often presented as a solution to adaptive feature extraction. However, transforming input signals into statistically efficient representations usually cannot produce meaningful high-level representations due to combinatorial explosion: there are simply too many potentially interesting high-level representations, most of which do not meet the requirements posed by the behavioural tasks at hand. Supervisory signals conveying information about the task requirements are therefore needed.

During evolution, some task requirements have become imprinted in genetically determined predispositions for developing certain types of representations. This alone is not enough for selecting suitable representations. Even if there could be enough information in the genes, evolution cannot hard-code the representations because environment and behavioural needs change from one generation to the next.

It is known from developmental psychology that active movement shapes perception considerably. In this article I propose an architecture in which sensorimotor coordination and goals of the behaving system shape perceptual representations to meet the requirements of behavioural tasks. The key adaptive component is a feature extractor where context-dependent expectations guide the development of feature extraction stage. I have earlier demonstrated that the developed features carry information about bottom-up primary inputs but which information is retained depends on the context-dependent expectations [1].

When connected into a suitable hierarchy, the context-guided feature extractors should be able to develop representations that are useful for controlling behaviour. A general rule for constructing the hierarchy could be that those systems that need the representations provide the contex or supervisory signals which guide the development of feature extraction. I will discuss predictive motor control and the development of affordances in detail but the arguments and architecture generalise to other behaviourally important attributes such as valence[1].

The rest of the paper is organised as follows. The next sections discuss the algorithmic (Sec. II) and biological (Sec. III) basis of developing invariant representations. Section IV proposes an architecture where representations of affordances develop through sensorimotor coordination and interaction with the environment. Extensions and future research are discussed in Sec. V.

II. Models of invariant feature extraction

Every child knows the concept of a “chair” and is apparently able to recognise chairs and use this information for guiding behaviour. We learn and recognise such behaviourally useful information so effortlessly that it is frustrating and embarrassing how difficult it is to build a computer vision system which could recognise chairs or a robot that could use this information.

The main problem with such categories is that their recognition requires complex nonlinear mappings from sensory information. As classical artificial intelligent research found out the hard way, it is extremely difficult to give useful formal definitions of most concepts and categories we have. Attention was therefore turned to neural networks which can *learn* things from experience.

[1]Affordance and valence refer to what can be done with something and how rewarding it is, respectively. They are qualities of objects in the same way as colour, shape, texture or timbre.

In machine learning, three main approaches can be distinguished: unsupervised, supervised and reinforcement learning. All of them have been tried for learning concepts but success has been limited. Supervised learning requires teaching signals which are clearly not available for developing humans or autonomous robots, and it takes extremely long time to learn complex nonlinear mappings. Reinforcement learning only requires reward feedback and such innate emotional responses are available. Unfortunately perceptual learning based on reward signals takes even more time than in supervised learning. Unsupervised learning aims at capturing whatever structure in input data and there has, indeed, been some success in learning simple categories. Mostly the data that has been used has still been very heavily preprosessed and learning categories like "chairs" autonomously from raw sensory data is still far from reality.

The unsupervised learning methods that have come closest to achieving the goal are based on finding *invariances*. An invariant pattern can take many forms but still be in some respects the same. Invariance is usually defined with respect to some transformation. Something is translation invariant if it remains the same in translation. A chair is a chair even if it is translated, scaled, rotated, illuminated differently, made of different materials and in different shapes. However, there is something that stays (roughly) invariant: one can sit on a chair. Usually these systems have been organised in a hierarchy, reminiscent to that found in the neocortex, with increasing invariance towards higher levels.

In order to extract invariant information, it is important to *discard* most useless information but simultaneously retain the essential information. Since the useful information can only be extracted by complex nonlinear transformations, there is a vast number of *potential* transformations and only a vanishingly small fraction of them results in anything like behaviourally useful categories.

The problem at hand can be defined as follows: 1) find a method that can select a nonlinear transformation which retains the desired type of information and 2) find an architecture where the supervisory signal for this selection can be made available. The remainder of this section addresses the first question.

A. Denoising source separation

Denoising source separation (DSS) [2] is a framework for developing source separation algorithms. It stems from research on principal and independent component analysis (PCA and ICA) [3] which fall in the category of unsupervised learning. Source separation methods are often described in signal-processing context as algorithms for extracting a source signal or a set of source signals from a mixture of signals. Since complex environments typically generate plenty of data and structured patterns, many of these methods are also useful as feature extraction algorithms. In other words, the distinction between signal separation and feature extraction is rather a question of the input data than the methods themselves.

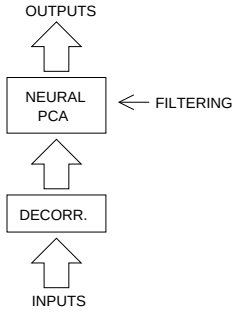

Fig. 1. A scematic illustration of a DSS algorithm. Learning is here implemented by neural PCA which uses a simple Hebbian rule but could be implemented by other numerical techniques such as power method. Filtering modulates the neural activations. Because the inputs are decorrelated, even weak modulation by filtering is enough to determine which type of information the system learns to extract.

The basic ingredients of DSS are decorrelation of inputs followed by filtering (denoising) embedded in iterative PCA (Fig. 1). Normally PCA would extract the component which has the highest variance but decorrelation removes all covariance structure of the inputs. This stage, called whitening or sphering in ICA literature, gives a decisive role for filtering in determining what type of information the system learns to extract. The amount by which filtering changes the outputs is not critical, only the *direction* matters. It is also possible to modulate the learning rate instead of the outputs, but filtering noise away from the outputs is often useful as such.

Depending on the selected filtering, DSS can implement various source separation algorithms ranging from (almost) fully unsupervised ICA to very specific, supervised learning algorithms [2]. What makes DSS interesting for the present discussion is its ability to get rid of unwanted information. This is related to so-called deflation methods in ICA [3] which can extract independent components one by one. This is in contrasts to many other unsupervised learning methods—generative models in particular—which typically need to model all structure in the data.

The basic DSS is a linear method and is therefore not suited for building complex nonlinear mappings. It is, however, easy to add a nonlinear feature expansion in the decorrelation stage [1]. These features can also be sensitive to various temporal aspects of the inputs. It is interesting to note that the properties of neocortical layer 4 network seem well suited for this task: feedforward and feedback inhibition acting on different temporal scales [4] can decorrelate and normalise the activations in space and time. As shown in [1], restricting the outputs to be positive is enough to provide suitable nonlinear features from which invariant representations can be built.

DSS with nonlinear feature expansion is certainly not the

only system that can learn to extract the desired information but it is simple, robust and computationally efficient.

B. Expectation-guided learning

The more difficult part of the problem is where the supervisory signal for selecting useful information comes from. Many existing unsupervised invariant feature extraction methods are based on temporal invariance [5]–[7]: it is a reasonable assumption that the identity of objects—such as the identity of a person—changes slower than sensory input—such as image on retina and sound to cochlea. Unfortunately slow temporal evolution does not guarantee behavioural significance. There definately seems to be a correlation between the two and many interesting invariant features can be learned from suitable data. The models of invariant feature learning have often used visual inputs as data and it may be that slowness is a relatively good criterion for the first stages of feature extraction in the visual system.

Unfortunately it does not seem possible to use slowness as a criterion for finding useful high-level concepts—at least nobody has yet provided compelling results despite several attempts. There are also cases where even the first stages of feature extraction cannot be based on slowness. This seems to depend on the nature of the inputs. Phonetic categories for examples change quickly and are invariant to some aspects of context but not invariant in time.

I have proposed [1] that a better alternative is to use learned expectations as the source of supervisory signals (Fig. 2). We need to assume that there is a context of some kind—this could include delayed, top-down or lateral inputs from other parts of the system—and the expectations about the outputs are composed from these signals. There would thus be two types of inputs to the feature extraction element: driving inputs from which the output features are composed and contextual inputs from which expectations about the output features are generated.

Since the driving inputs are strong, the output activations reflect primary inputs. For the same reason, the development of expectations is guided by primary inputs. However, since the primary inputs are whitened, Hebbian, correlation based learning does not prefer any direction in the input space. Therefore a modulatory influence of the expectations can guide learning and determine which piece of bottom-up information the system learns to represent. This is a kind of matching process which selects the coherent parts of bottom-up and contextual inputs.

In this framework, methods which rely on temporal invariance can be interpreted to use past outputs as expectations of the future outputs. The system thus tries to find any features that remain constant over time. In the architecture depicted in Fig. 1, this can be implemented by simple low-pass filtering.

If the expectations are not simply delayed outputs but are learned mappings from the past outputs, the system

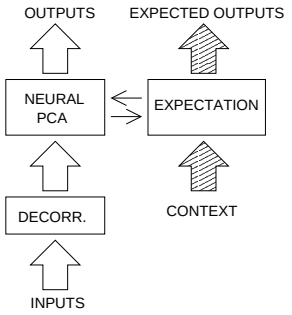

Fig. 2. Expectations can be used for guiding learning in DSS. The driving input dominates the activations over the context and also drives learning of the expectations. Due to decorrelation of the driving inputs, the expectations composed from the context can still drive learning of the output features.

can learn to extract features with structured spectral characteristics or—if nonlinear predictions are allowed—even features exhibiting complex nonlinear dynamics [8]. Note that the extracted features do not need to change slowly: for instance rapidly changing phonetic categories could be learned as long as the context provides information from which expectation about the phoneme can be composed. If the context changes rapidly, the target category can change rapidly.

C. Modelling results with different types of contextual information

Expectation-driven feature extraction offers a versatile framework where the type of information extracted from the inputs can be controlled by selecting suitable contextual information. At least three types of contextual information can be distinguished: temporal (or delayed), lateral and top-down context.

Temporal context refers to delayed outputs of the feature-extraction module itself and therefore corresponds to unsupervised learning. Nonlinear dynamics was shown to extract and separate three nonlinearly mixed dynamical processes in [8]. The method used a nonlinear generative model and therefore is unable to select information: all dynamical processes need to be extracted simultaneously. The same principle can be used with DSS and then the processes can be extracted one-by-one at least from linear mixtures (Alexander Ilin, unpublished results). In [9] it was shown that DSS with temporal context is able to select structured features from high-dimensional noisy data: El-Niño Southern Oscillation and several other structured weather phenomena were extracted from a large climate dataset with 30,000-dimensional daily measurements over more than five decades. Minimal assumptions about the structure of the desired features were enough for the extraction.

Of all the cortical areas, the primary visual areas have been studied most extensively. Many models of invariant feature extraction have therefore been tested with simple

visual stimuli and have been shown to develop simple edge filters and a subsequent transformation to invariant edge filters, similar to the so-called simple and complex cells in the primary visual cortex V1. Such invariant features emerge as the features exhibiting slow temporal evolution when trained with sequences of natural images (e.g., using adaptive-subspace self-organising map as in [6] or slow feature analysis as in [7]).

This has sometimes been taken to mean that the sensory system uses slowness as the criterion for developing invariant features but in [1] it was shown that expectations derived from lateral information will also induce the development of similar translation invariant edge detectors. Lateral information refers to the activations of adjacent areas, in this case adjacent image locations. An edge in one image location predicts a continuing edge in nearby areas, but the exact spatial location may not be well determined. This uncertainty in the prediction promotes the development of invariant features in the expectation-guided learning architecture depicted in Fig. 2. In [1], the lateral context consisted of the outputs of neighbouring modules.

An efficient design for feature extraction architectures is a pyramidal hierarchy where each module processes inputs from a few modules at the lower level. Such a hierarchy offers a third type of information for composing the expectations: top-down context. In such hierarchies, top-down signals have necessarily less resolution than the bottom-up features simply because by design, there are less representing elements at the higher levels. Again, this imprecision of expectations can drive the development of invariant features.

To summarise this discussion about the algorithmic basis of developing invariances, expectations derived from contextual information can drive the development of invariant features which reflect the predictable structure in the observations. Since the architecture shown in Fig. 2 learns to extract any information that can be predicted, the type of information extracted from the inputs can be controlled by selecting the information included in the context and also by restricting the capabilities of the system producing the expectations. Note that the contextual input conveys supervisory signals for selecting suitable information from the inputs but the supervisory signals need not be explicit. The bottom-up inputs also select those parts of the context that are suitable predictors of the inputs.

III. Context selection as a means of guiding learning

The previous section identified a method that can select a nonlinear transformation which retains the desired type of information. It turned out that many types of contextual information can supervise the development of this selection but we still need to find an architecture where a suitable context can be made available. In this section I propose that the brain uses two complementary mechanisms for providing the context: 1) genetically determined reciprocal connections between areas and 2) goal-directed attentional filtering. The first is a relatively static, genetically determined mechanism while the latter relies on dynamic, goal-directed routing of relevant information between cortical areas (covert attention) and orientation towards relevant stimuli (overt attention).

A. Behavioural requirements

Let us first consider what kind of representations are needed for controlling behaviour. Coordination of motor output (and to a lesser extent, hormonal output) is the only thing a brain is needed for. The motor output is, in turn, a means for achieving goals and therefore the brain needs to invert desirable end results into their causes, muscle movements. Furthermore, behavioural control occurs over many timescales ranging from immediate motor control to planning one's life.

Sensorimotor transformations are most important for immediate motor control while longer-term planning requires prediction of future events and in particular consequences of actions. Furthermore, the success of planning depends crucially on accurate evaluation of the consequent situations and therefore it is the emotionally significant consequences of actions that should be predicted. In summary, those things are worth representing which are able to predict actions and emotions.

B. Interplay between specialised structures and the neocortex

With this background, it is not surprising that in the brain, most things are somehow related to actions or emotions. Most sub-cortical areas are closely involved in behavioural output, evaluation or both. These areas are often heavily interlinked with the cortex and it is commonly considered that during mammalian evolution, neocortex has taken over many of the functions of the specialised systems—this is referred to as corticalisation. In humans this process has gone furthest and we have, for instance, more control over our spinal motoneurons and emotional structures than other mammals.

When considering the behavioural timescale, it seems obvious that neocortex is running the show. However, the situation looks somewhat different on learning timescale. I propose that sub-cortical structures play a major role in providing inputs for the cortex. This input guides learning in the cortex and guarantees that the cortex provides information which is useful for the sub-cortical structures.

There are many sub-cortical structures that are reciprocally connected with those areas of the neocortex that process the same type of information. Partly this is certainly because the relevant part of cortex receives the right kind of sensory information but I propose that the outputs of sub-cortical structures are used as context which guides learning and guaranteeing that the right type of information is extracted from the inputs. In this context, "right type" means

information that the corresponding sub-cortical structure can use.

C. Cortico-cortical context

While the sub-cortical and peripheral sensory signals may define the "extremal points" of sensori-moto-emotional associations, the cortex seems to be designed to find the missing links: representations, concepts and categories that link different sensory modalities with emotions, movements and other such "internal" modalities.

Neocortical areas are heavily interlinked. The first impression from diagrams of cortical connectivity is that everything is connected to everything else. While the connectivity is more specific than that, it is nevertheless clear that any cortical area receives connections from many other areas. However, a large proportion of of these connections are "modulatory"(e.g., top-down connections fall in this class) which means that they have relatively weak influence on the activations. The thalamo-cortical bottom-up connections to layer 4 are quite specific and although they represent a relatively small portion of connections to any cortical area, they functionally much stronger than most other inputs.

In light of expectation-guided learning (Fig. 2), this suggests that each cortical area is learning to extracting information from the bottom-up stimuli under the guidance of contextual inputs from other cortical areas. This would mean, for instance, that the visual system could learn visual categories which would reflect distinctions in auditory stimuli. This would be useful for example in learning lip-reading. In general, heavily interconnected areas would tend to learn coherent representations of the environment.

D. Attentional filtering mediating goal-directed information

The above cases suggest that evolution has set up a network of connections mediating contextual information that can guide learning. However, it is unlikely that such fixed contexts could be enough for learning complex representations and abstract concepts. The problem is, again, that there is too much structure in the environment. It seems likely that the cortico-cortical connections would mediate too much information to be useful in guiding learning. The information needed for guiding learning in a given cortical area could be available but there would be a large amount of other, irrelevant information which would interfere with learning.

Attentional filtering is a good candidate for a process which could select the right type of contextual information. Attentional filtering is a competitive process which selects information such that only some information in the sensory stimuli reaches higher levels. Moreover, attention has a strong goal-directed component. These features make attentional filtering a promising candidate for guiding learning.

It is, in fact, well-known in psychophysics that attention has a major effect on learning. Some researchers have even questioned whether there is any perceptual learning without attention. While attention is considered important in learning, it has been less clear what exactly is the underlying mechanism and principle. I propose that attentional filtering selects the content of the contextual information that cortical areas receive, the context determines what kind of expectations are generated and this in turn guides learning.

I am not aware of conclusive data supporting this hypothesis but it is know that the development of interactions between areas depends on attention [10] and attention and task-context has a strong influence on learning perceptual categories [11].

E. Biased competition model of attention

Both learning invariances and attentional filtering are processes which select useful information. The main difference is that they operate on different timescales. What development of invariant features does during learning, attentional filtering does on behavioural timescale. Interestingly, the neurodynamical model of attention [12], [13] shares many characteristics of the architecture proposed for invariant feature extraction in Sec. II: hierarchy of processing areas where top-down expectations bias the competing activations. It should be noted that while competition is not explicit in Figs. 1 and 2, it is an important part of neural PCA.

As a result of local competition, biasing by expectations from long-range contextual input and the pyramidal structure of the processing hierarchy, an attentional process emerges: only some sensory stimuli have access to high-level processing areas. This emergent selection process has both top-down and bottom-up components. Salient bottom-up stimuli tend to win the competition which is nevertheless biased by top-down expectations.

F. From specific to invariant and back

While attention and learning invariances thus seem to rely on very similar mechanisms and achieve selection of information, they complement each other in an interesting way. A forward transformation from specific input patterns to invariant features develops gradually during learning because decorrelation of the bottom-up inputs gives a decisive role for the top-down driven expectations even if they only modulate the activations very weakly as discussed in Sec. II.

As long as the activations are predominantly driven by bottom-up inputs while still being modulated by top-down context, the top-down information flow implements the inverse transformation from invariant features to specific patterns. Information at the highest levels thus percolates down the hierarchy as long as the attentional process has a functional top-down component. This can happen only if the top-down expectations have a clearly non-zero biasing effect on the local competition between activations.

The biasing effect can be made more robust by bottom-up normalisation which operates on the competing neural assemblies and emphasises the activations of neurons belonging to weaker assemblies over the activations of neurons belonging to stronger assemblies. The result is that top-down biasing has a more important role in deciding the outcome of local competition even if it is weak. As discussed in [1], there is evidence of such normalisation in cortical representations.

G. Expectations and associations

The attentional process can distribute high-level supervisory signals, such as working memory contents reflecting the goals of the system, to lower levels, but it cannot propagate them in time. This is where predictive associations come to rescue. Recall that the representation should help predicting actions and emotions. It is therefore not surprising that the brain has mechanisms for predictively simulating real world. These mechanisms underlie our ability to imagine and expect stimuli. The predictive expectations accompanying perception have been studied for instance in psychophysical priming experiments. The predictions are useful as such but they also mean that the contextual information reflects future actions, emotions and sensory percepts, that is, the context promotes the development of representations which are useful for prediction.

IV. Architecture for affordances: how reflexes could teach vision

In this section I will outline a story of how the mechanisms proposed in the previous section could cooperate in learning affordances.

Many developmental psychologists emphasise the role of activity in learning. A dramatic example is provided by the so-called kitten-carousel experiment [14] where two kittens are raised in similar visual environments but only one of the kittens can move. Its movements are mechanically mirrored to the other kitten which therefore receives the same visual stimuli, the only difference being that the passive kitten did not itself create the movements which induced the visual stimuli. The result is that only the active kitten develops functional depth vision.

Based on behavioural experiments alone it is not possible to know whether the deficit is caused by lack of integration between vision and action or by deficit in early visual processing. It is possible that the passive kitten perceives depth but is unable to incorporate the information into behaviour and therefore keeps bumping into objects. However, there is a lot of other evidence supporting the important role of attentive, active behaviour in preceptual learning [11].

A. Predictive motor control

It may sound a bit strange at first that reflexes (or more properly innate behaviours) could teach movements but reflexes are actually a very natural way of encoding desired behaviours.

In predictive control, corrective reflexes teach a predictive controller to make anticipatory movements (e.g., [15]). One of the best understood systems is gaze stabilisation where a hard-wired, innate optokinetic reflex (OKR) elicits eye movements which tend to counteract optic flow detected on retina. Importantly, OKR also provides the target for learning anticipatory eye movements, such as the vestibulo-ocular reflex (VOR) which uses the signals from the balance organ to elicit compensatory eye movements even before the retinal slip has been detected. (See [16] for a review of details and computational models of OKR and VOR.)

VOR is adaptive as has been demonstrated in prism experiments which completely reverse the direction where eyes should move in order to compensate head movements. A key element in VOR appears to be the prediction provided by the cerebellum. Moreover, OKR/VOR is just one example of predictive control and cerebellum is more generally involved in accurately timed prediction tasks.

B. Proposed architecture

In order to fulfill its prediction tasks, the cerebellum needs suitable sensory representations. Since the cerebellum tries to predict, the predictions themselves contain the best contextual information to supervise the development of representations that the cerebellum can use for compiling its predictions. Consider for instance the grasping reflex. When an object is placed in the palm of an infant, she will reflexively grasp the object. Over time, the infant learns to anticipate the reflex, for instance from visual information that the cerebellum receives. Initially the prediction is crude but still sufficient to provide the correct direction. Graspable objects elicit a prediction of the motor act of grasping.

This cerebellar prediction is well timed unlike the cortical associations which run forward without respecting real time. Nevertheless, it underlies sensorimotor coordination which creates structured sensorimotor patterns. Their motor aspects are represented by the cortical motor hierarchy which receives motor information from the muscles and predictions from the cerebellum. It may be that visual context is important for teaching the motor system an invariant representation for grasping: this sequence of movements occurs in this visual context. What is clearly important is that the motor context of grasping teaches the visual system to extract information related to movements: this object is graspable.

Figure 3 illustrates the proposed developmental scheme. Reflexes are the primary source of supervisory signals. They teach the cerebellum to predict movements such as grasping. Initially the prediction can be based on crude visual information. The initial sensorimotor coordination with the environment generates structured sensory percepts (including motor percepts) which are represented by motor, visual and other sensory hierarchies. Cortical associations

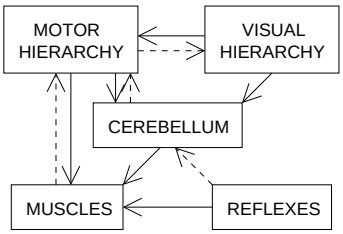

Fig. 3. Information flow in the proposed architecture for learning visual affordances. Solid arrows denote forward flow of information and the dashed arrows denote the flow of supervisory signals. Note that in some cases the distinction is functional, not anatomical: in motor cortex, the anatomical flow of information is from elementary muscle movements towards high-level representations. Also, this diagram only concerns with the information flow inside the learning system. System-environment interaction determines which reflexes are triggered and what visual information the system receives.

and attentional process distribute the contextual information forward in time and down the cortical hierarchies, providing the target for learning affordances and other behaviourally relevant features.

V. Discussion

The above story concentrated on the development of affordanced guided by innate reflexes. Similar considerations generalise to other innate systems such as the value systems and reinforcement learning. There are innate value systems which signal rewarding stimuli, such as the taste of sugar. Other systems turn these signals into predictions of future reward which are believed to be conveyed by mid-brain dopaminergic projections. These predictions and other related emotional signals are represented by certain temporal and frontal cortical regions and guide the extraction of valence in sensory systems. Reward predictions are also used by the motor hierarchy for reinforcement learning of rewarding sensorimotor associations, thus providing another source of innate movement-related signals, and control of attention which gates the access of sensory information to higher levels of processing and context.

In general, it seems that many sub-cortical structures have reciprocal connections with certain cortical areas. For instance, it is interesting to note that superior colliculus, a structure dedicated for orientating behaviours, has reciprocal connections with cortical areas which are part of the "what" pathway. Such connection patterns may reflect the need to make sure that the cortical area receives suitable contextual information to guide the development of features which the sub-cortical structure needs.

Many of the ingredients of the proposed architecture have been modelled in isolation and in general the architecture agrees well with many known features of the brain. Furthermore, many of the ingredients required by the complete architecture—most notably attentional process and predictions—are useful as such. Nevertheless, there is obviously still hard work to be done and many things that could be incorporated in the model.

A. Conclusion

Sensory stimuli of behaving systems have a potentially large number of invariant structures, only some of which are behaviourally relevant. In this article I proposed algorithms and architectures to extract suitable representations in a semi-supervised manner. Sensory data largely determines which types of representations are potentially useful. Structures involved in generating and evaluating behaviour also generate intrinsic contextual signals which guide the selection among suitable representations. The general rule is that a specialised system can provide its own outputs as contextual information for the cortex. The context is distributed forward in time and over the cortical hierarchy by associations and attentional process. The expectations derived from the contex then guide the development of representations which the specialised system needs.

Acknowledgements

This work was partially funded by European Commission project ADAPT (IST-2001-37173).

References

[1] H. Valpola, "Behaviourally meaningful representations from normalisation and context-guided denoising," Artificial Intelligence Laboratory, University of Zurich, Tech. Rep., 2004, available at Cogprints http://cogprints.ecs.soton.ac.uk/archive/00003633/.

[2] J. Särelä and H. Valpola, "Denoising source separation," *Journal of Machine Learning Research*, 2005, in press.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. Wiley, 2001.

[4] M. Beierlein, J. R. Gibson, and B. W. Connors, "Two dynamically distinct inhibitory networks in layer 4 of the neocortex," *Journal of Neurophysiology*, vol. 90, pp. 2987–3000, 2003.

[5] P. Földiák, "Learning invariance from transformation sequences," *Neural Computation*, vol. 3, pp. 194–200, 1991.

[6] T. Kohonen, S. Kaski, and H. Lappalainen, "Self-organized formation of various invariant-feature filters in the Adaptive-Subspace SOM," *Neural Computation*, vol. 9, no. 6, pp. 1321–1344, 1997.

[7] L. Wiskott and T. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural computation*, vol. 14, pp. 715–770, 2002.

[8] H. Valpola and J. Karhunen, "An unsupervised ensemble learning method for nonlinear dynamic state-space models," *Neural Computation*, vol. 14, no. 11, pp. 2647–2692, 2002.

[9] A. Ilin, H. Valpola, and E. Oja, "Semiblind source separation of climate data detects El-Niño as the component with the highest interannual variability," in *Proc. Int. Joint Conference on Neural Networks (IJCNN'05)*, Montréal, Canada, 2005, submitted.

[10] W. H. R. Miltner, C. Braun, M. Arnold, H. Witte, and E. Taub, "Coherence of gamma-band eeg activity as a basis for associative learning," *Nature*, vol. 397, no. 6718, pp. 434–436, 1999.

[11] C. D. Gilbert, M. Sigman, , and R. E. Crist, "The neural basis of perceptual learning," *Neuron*, vol. 31, pp. 681–697, 2001.

[12] G. Deco and B. Schürmann, "A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition," *Vision research*, vol. 40, pp. 2845–2859, 2000.

[13] G. Deco and E. T. Rolls, "A neurodynamical cortical model of visual attention and invariant object recognition," *Vision research*, vol. 44, pp. 621–642, 2004.

[14] R. Held and A. Hein, "Movement-produced stimulation in the development of visually guided behavior," *Journal of Comparative and Physiological Psychology*, vol. 56, no. 5, pp. 872–876, 1963.

[15] M. Kawato, "Feedback-error-learning neural network for supervised motor learning," in *Advanced Neural Computers*, R. Eckmiller, Ed. Elsevier, North Holland, 1990, pp. 365–372.

[16] M. Kawato and H. Gomi, "The cerebellum and VOR/OKR learning models," *Trends in Neuroscience*, vol. 15, no. 11, pp. 445–53, 1992.