Denoising source separation: a novel approach to ICA and feature extraction using denoising and Hebbian learning

Jaakko Särelä[1] and Harri Valpola[2]

[1] Neural Network Research Centre, Helsinki University of Technology
P.O.Box 5400, FI-02015 HUT FINLAND
[2] Laboratory of Computational Engineering, Helsinki University of Technology
P.O.Box 9203, FI-02015 HUT, FINLAND

Abstract

In this paper, we review the recently proposed denoising source separation (DSS) framework. In the DSS framework, source separation algorithms are constructed around denoising proceduces. The denoising should reflect the prior knowledge of the source characteristics and it can be procedural. Source separation methods are an active research topic in signal processing domain but they can also be applied to feature extraction. It has also been proposed that independent component analysis (ICA) and related methods have similarities with sensory processing in the brain. The main purpose of this paper is to discuss extensions to the basic DSS framework to make it more suited for feature extraction. We also discuss connections with sensory processing in the brain.

1 Introduction

The goal of feature extraction is to find a representation that facilitates the task at hand, such as object recognition or motor control. Obviously the features ought to capture the important information in the incoming data but the information-theoretic notion of information as something that can be decoded is not the same as something that can be decoded in practice. Useful information should be made as explicit as possible so that it is easy to decode it. In addition, useless information should be discarded so that it is easy to learn to use the useful information.

The traditional principal component analysis (PCA) extracts features which explain the largest amount of variance in the data that is possible with a linear mapping. PCA can be realised by simple Hebbian learning, for example, by Oja's rule [Oja, 1982] and has attracted interest as an abstract model of cortical sensory processing. However, the features provided by PCA appear very different from those found in the brain. Moreover, PCA tends to emphasize correlations between sensory inputs while in the brain, mechanisms that *remove* correlations abound.

Independent component analysis (ICA) [Hyvärinen et al., 2001] has been more successful in developing features which resemble those found in the brain, at least in the primary visual cortex (simple cells in V1). While the features are only guaranteed to be uncorrelated in PCA, ICA renders the sources as independent as possible. This seems to be a reasonable criterion for making information explicit: one feature encodes one independent aspect of the environment. Another, related criterion is sparsity of the features. Again, this appears to be a reasonable criterion for feature extraction because different entities in the environment can often be accepted to appear infrequently. Most of the time the feature would be absent while sometimes it would be strongly activated. This amounts to a sparse temporal distribution. Often, but not always, ICA and sparse coding develop similar features (see, e.g., [Olshausen and Field, 1996] for sparse coding and [Hoyer and Hyvärinen, 2000] for independent component analysis of natural images).

Basic ICA and sparse coding rely on linear mappings and it is quite difficult to extend them for nonlinear mappings. This is a serious limitation for feature extraction because often the important features in the data can be made explicit only by nonlinear transformations. So-called subspace methods have had some success in developing hierarchies of increasingly nonlinear features. The idea is to group together several simple features that appear to encode the same entity. This creates a "complex" feature which can be more *invariant* than the simple features. The simple features can respond, for instance, to rotations, translations or scalings of the same object. The combined complex feature would then be invariant to rotation, translation, scaling or all of them. The most common criterion for grouping simple features is temporal proximity of the activations (see, e.g., [Kohonen et al., 1997, Wiskott and Sejnowski, 2002, Hyvärinen et al., 2003] for models which develop features akin to simple and complex cells in V1). It is reasonable to assume that features that are often activated at roughly the same times, signal the presence of the same entity.

None of the criteria discussed above (high variance, independence, sparsity, temporal proximity) is ecological in the sense that it would take into account the subsequent use of the features for tasks such as recognition or control. This becomes a serious limitation when deeper nonlinear hierarchies are built because the number of *potential* groupings and nonlinear features grows quickly. In deeper hierarchies it becomes crucial to select and guide the development of feature extraction by signals which carry information about the goals.

We have recently introduced the denoising-source-separation framework [Särelä and Valpola, 2005] where separation algorithms are constructed around denoising procedures. Previously we have mainly focused on signal processing and data analysis. We review some of that work but mainly concentrate on how DSS can be applied to goal-driven learning of nonlinear features. In DSS, the information about the goal is presented in a form of a denoising function. In this paper we focus on expectation-driven learning where contextual information is used to predict and denoise the features. This promotes the development of features that *can* be predicted. We report results from nonlinear feature extraction, discuss how this may be related to sensory processing in the neocortex and present a hypothesis about how goal-related information can be propagated in hierarchical models.

2 Denoising source separation

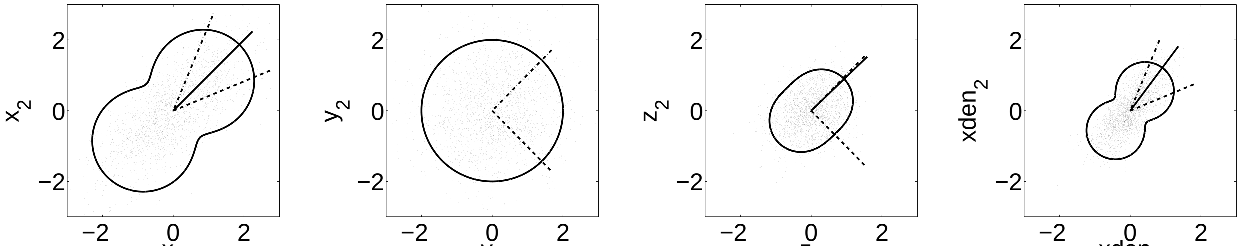
Source separation methods are often described in signal-processing context as algorithms for extracting a source signal or a set of source signals from a mixture of signals. Since complex environments typically generate plenty of data and structured patterns, many of these methods are also useful as 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.

In this section, we introduce DSS and explain how it is able to perform linear source separation and feature extraction. We also demonstrate the use of DSS in a computationally demanding problem with real-life climatological data.

2.1 DSS in source separation

Consider a linear mixing of independent sources. Each source causes correlations between the different mixture components and the total correlations are the sums of the individual correlations. Hebbian learning can be used to identify the direction with maximal variance. This can be implemented neurally [see for instance, Oja, 1982] but classical statistical methods such as the power method are available, as well. In order to learn a sub-space or maximal variances, anti-Hebbian learning may be used to keep the different directions separate. This induces a *competition* between the projections and leads to uncorrelation of the projected components.

Consider the two-dimensional input in Fig. 1a where two independent sources have been mixed. The dashed lines indicate the mixing directions. Some clear correlations exist between the dimensions and this becomes visible in the nut-shaped curve that describes the variances of the projections in each of the directions. In order to identify a single source, one needs to find the projection that cancels the contribution of all the other sources. This would happen, when the projection is orthogonal to all the other projections. However, the largest variance direction, identifiable using Hebbian learning, does not usually identify any of the sources, but a mixture of them.



(a) (b) (c) (d)

Figure 1: *(a) Scatter-plot of the mixtures, the nut-shaped curve indicates standard deviations in each projection direction and the solid line the direction of the largest variance. The dashed lines indicate the mixing projections of the original sources. (b) Whitened data. (c) Effect of denoising of the whitened data. (d) Effect of the denoising of the original data.*

Assume now that the source corresponding to the dash-dot line is known and the corresponding projection should be learned. Learning of the projection converges to the dash-dot line, but the projection does not identify the source. This is because the other sources contribution has not been cancelled. Thus it seems more important to be able cancel the contribution of the other sources than to identify the projection of the needed source. But this would lead to a situation where all the sources should be known.

On the other hand, should the mixing be orthogonal, the identification of the projection would lead to separation. The orthogonality can be achieved by whitening (decorrelation and normalisation of variances). The result of this operation is shown in Fig. 1b. Knowing of the source would now result in finding the projection that identifies it. However, this requires that the source is known completely.

Assume instead that there is some known structure in the source (for example temporal) that is enhanced by a filtering (or denoising) operation. This is illustrated in Fig. 1c). The variance of the whitened data is unity in all directions. However, the denoising operation changes the variance structure, The denoising operation usually loses information of the desired source. However, it is only needed that the operation cancels even more the other sources. In that case, the desired source can be identified by another Hebbian learning stage, as indicated by the alignment of the solid line and the dash-dot line in Fig. 1c). Thus the whitening made the data sensitive to subtle operations such as the denoising. If no whitening had been used before the denoising, the results would have been almost identical to simple Hebbian learning without any denoising. This is illustrated in Fig. 1d)

The above source separation scheme had three stages: 1) decorrelation of the inputs and normalisation of their variances (whitening, anti-Hebbian learning), 2) denoising and 3) Hebbian learning

(PCA). We call such a procedure denoising source separation (DSS). The second and third stages are separate only if the denoising is linear that is it can be realised by matrix multiplication. If the denoising is nonlinear, the latter stages are molded into one, and DSS resembles nonlinear PCA [Oja et al., 1991]. The third stage has to be accompanied by some kind of anti-Hebbian learning as well, otherwise the different sources converge to the same solution. In the first stage, the anti-Hebbian learning made the following sensitive to subtle operations (denoising). DSS was first realised as a batch algorithm [Särelä and Valpola, 2005], but by using neural Hebbian and anti-Hebbian methods accompanied with online denoising, it can be implemented neurally as well.

2.2 DSS in climatology

In this section, we review one application of DSS in climatology. The data are dense-grid global measurements of surface temperature, sea level pressure and precipitation during several tens of years, resulting in a large dataset with tens of thousands of dimensions and thousands of time points. The aim is to demonstrate the capacity of DSS to adapt to the tasks at hand and to find the important information from the large dataset containing myriad structures. To be able to deal with large datasets, we have as well developed several speed-up and stabilisation methods [see especially Valpola and Särelä, 2004].

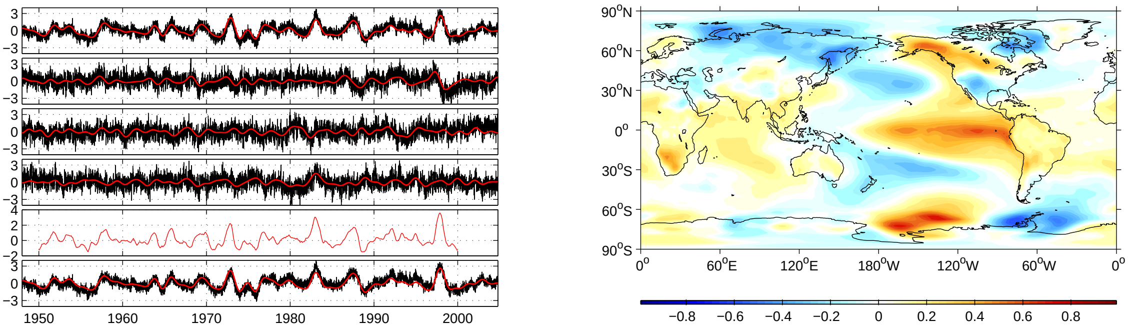
El Niño Southern Oscillation (ENSO) is a global-scale phenomenon in the ocean and atmosphere, known as one of the most prominent sources of interannual variability in weather and climate around the world. Its oceanic component (El Niño—La Niña events) can be defined as a Pacific basin-wide increase (El Niño) or decrease (La Niña) in the sea surface temperatures in the central and/or eastern Pacific Ocean [Glantz, 1996]. The warm El Niño events are known to be accompanied by the decrease in the sea level atmospheric pressure in the western Pacific, which is the atmospheric component of ENSO called Southern Oscillation (SO). SO can be defined as a large-scale oscillation of the air mass between the southeastern tropical Pacific and the Australian-Indonesian regions.

Ilin et al. [2005] searched for physically meaningful states with slow, interannual time course from climate data. Diurnal and annual variability in solar radiation means that the climate system has daily and annual cycles but climate events such as ENSO whose time course is slower than the annual cycle are in evidence that the climate system has intrinsic interannual dynamics. They showed that the basic ENSO events (El Niño and SO) appear as the component with most prominent interannual variability from the global measurements of surface temperature, sea level pressure and precipitation. Some of the results are reproduced in Fig. 2.

2.3 DSS in feature extraction

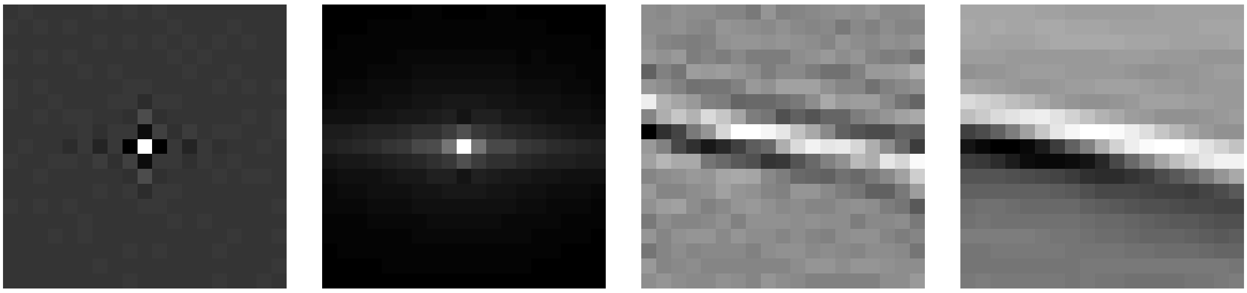
To exemplify feature extraction, we shall consider a set of images taken from natural scenes. Now the question becomes: what kind of features could be used to describe the content of the images. In the primary visual cortex (V1), simple cells have been shown to have Gabor-filter-like receptive fields [Olshausen and Field, 1996]. Such features would thus offer a good starting point.

First, let's consider the features achieved by simple (symmetric) anti-Hebbian learning or decorrelation. We sampled 19×19 patches from an image picturing a natural scene. One decorrelating projection is shown in Fig. 3a. The gray levels correspond to the forward projecting weights implementing decorrelation. The feature effectively takes the middlepoint and subtracts its neighbours to achieve decorrelation. This kind of feature resembles the on-center-off-surround cells of retina and thalamus. The inverse of the decorrelation filter is shown in Fig. 3b. This corresponds to the input image that activates only this feature.



(a) (b)

Figure 2: *(a) Climatological slow components. The first is caused by the El Niño effect. (b) The spatial map corresponding to the El Niño component.*



(a) (b) (c) (d)

Figure 3: *a) A simple feature from a natural image using symmetric decorrelation (decorrelation projection). b) The signal which activates the feature (a). (c) An edge-detector feature found by DSS (projection weights). (d) The input signal corresponding to (c).*

Combining the decorrelation with the other two DSS steps, denoising and PCA, produces structured features, one of which is shown in Fig. 3c. The feature responds to an edge across the image patch. This result is in concordance with using ICA as feature extraction technique and as a model of the simple cells in V1 (see, for instance [Olshausen and Field, 1996, Hyvärinen et al., 2003]). The corresponding signal which activates the feature is shown in Fig. 3d. Here a so-called shrinkage function was used for denoising. With this denoising function, DSS becomes close to ICA and sparse coding.

Now that we have discussed DSS as a source separation technique and a feature extraction technique, it becomes useful to summarise the method schematically. This is depicted in Fig. 4. As mentioned before, nonlinear denoising has to be put inside the PCA stage.

3 DSS and nonlinear separation

In the methods shown so far, denoising was nonlinear in the edge-filter example, but all examples used linear mappings to produce the features or extracted source signals. In order to create complex features, nonlinear mappings need to be employed. One possibility is to use feedforward networks with hidden layers that have nonlinear output mappings. Such networks are often used in machine-

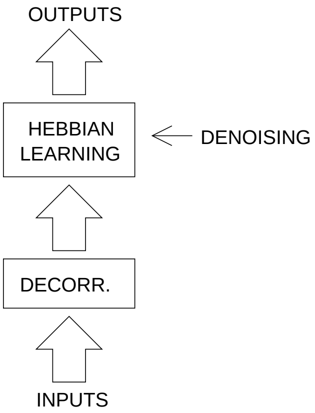


Figure 4: *A schematic illustration of DSS.*

learning community for supervised learning, such as classification and pattern recognition. In this section we show that multi-layer perceptron (MLP) networks can also be used in DSS but the applicability seems to restricted to low dimensions. In high-dimensional problems, methods related to nonlinear regression seem to be more reliable and efficient.

3.1 Nonlinear separation using an MLP network

Consider the two images in Fig. 5a. The data are two scanned 435x322 images. Before scanning, the images have been printed on two-sides onion-skin paper. Since onion-skin paper is transparent, this produces a nonlinear mixture (Fig. 5b) of the original images. Nonlinear ICA has been used to solve this problem [Almeida and Faria, 2004]. The images have been preprosessed with symmetric whitening, which tries to preserve the original dimensions as well as possible. For this reason the images are relatively well identifiable already (Fig. 5c).

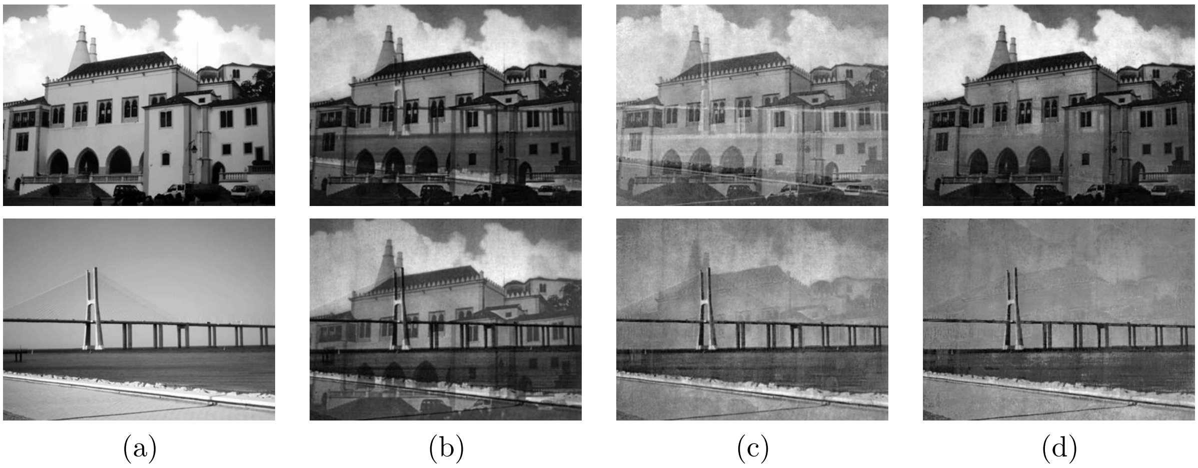


Figure 5: *(a) Original images. (b) The nonlinear onion-skin-paper mixtures. (b) Results after symmetric whitening. (d) Results using the MLP-DSS and wavelet denoising.*

As a first separation step, DSS with linear observation mapping was used. The denoising was as follows: The mixtures were transformed into a sparse wavelet basis with suitable detail. Then, in the wavelet space, a competition was introduced between the sources. No other decorrelation criteria were imposed on the sources. After convergence, the denoised output was used as the target to learn an MLP mapping. Finally, MLP-learning and denoising were alternated for 10 times. At each iteration the MLP learned until it converged. The results in Fig. 5d show that this procedure was able to learn a reasonable nonlinear separation mapping. However, learning is not very stable and with more iterations the results gradually degrade.

It is difficult to have nonlinear feed-forward networks to learn complex nonlinear mappings, even when there exists correct supervised output. This is because the model can learn almost any nonlinear mapping and thus overfits easily. Learning unsupervised MLP-DSS is even more difficult, because learning cannot be guided by a correct output, but instead one has to resort to imperfect denoising mechanisms.

3.2 Nonlinear feature expansion

In addition to already mentioned difficulties, the simple MLP approach to nonlinear mapping does not work well in large dimensions: the learning becomes extremely slow and overfitting even more probable.

Another possibility is to make some kind of a feature expansion before the Hebbian learning and the denoising. Such an approach have been used by Blaschke and Wiskott [2005] to get a nonlinear ICA algorithm, and comes close to nonlinear regression. Another nonlinear expansion was used by Valpola [2004]: a shrinkage function that was restricted to give positive outputs.

In this paper, we speculate about a general method for creating the needed nonlinearity. It is at least necessary to restrict oneself to only mild nonlinearities, especially in high dimensions. This restriction is needed to avoid overfitting, as well. The aim of the nonlinear feature expansion is to make it possible to learn complex features that are composed of simple features. This is more than what can be achieved by simple linear models.

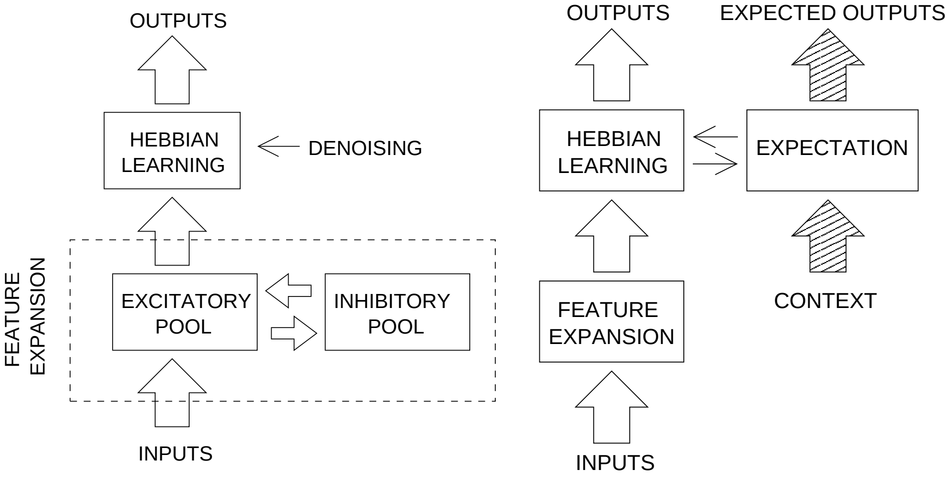
The computations on layer four in the neocortex resemble the idea of feature expansion (see e.g. Thomson and Bannister [2003], for a discussion of the connections between different layers in the neocortex). The computation is based on an interplay between pools of excitatory and inhibitory neurons. The inhibitory pool of neurons induces a competition between the excitatory neurons making only few of them active simultaneously. We propose that this method can be used for the nonlinear feature expansion in DSS. Additionally it is necessary to require that the output of the feature expansion is positive.

A schematic illustration of the whole DSS process with the nonlinear feature expansion is given in Fig. 6.

4 Hierarchical models

Packing a hierarchy of linear modules is a waist of time: layers of linear mapping stay linear. With the nonlinear feature expansion layer, hierarchical systems become powerful tools. Slow feature analysis [Wiskott and Sejnowski, 2002] is an example of such a hierarchical model. However, its problem is that in no part of the system, the learned features are guided by the task at hand. And this would ultimately be the goal of the feature extraction. For instance, in some cases there exist data from several modalities. Then it would be useful to extract features that integrate information between the modalities. Another, more general way of putting this would be to say that it is necessary that expectations guide the formation of the features.

In this section, we first suggest a hierarchical model where expectations are used to guide the learning. One experiment reported in Valpola [2004] is reviewed. The results show that the grouping



(a) (b)

Figure 6: *(a) DSS layer with the feature expansion, The pools of excitatory and inhibitory neurons are shown inside the dashed-line box. (b)A schematic illustration of context-guided learning in DSS.*

of simple features into more complex ones can be guided by expectations. Then, in Sec. 4.2, we speculate the role of attention in guiding learning by modulating the expectations.

4.1 Context-guided learning

Every learning situation has a context in which it happens. It is reasonable to assume that the context creates expectations of what is present in the data. As the context is part of the environment and thus part of the data itself, it is natural to use the features extracted in one part of the model as the context of the other parts and other time instants. The system can thus be symmetric, every part of the data being used as the context for the other parts. Because the different parts process different information, the context-guidance makes each part to learn new things. Though the learning of the features is guided, this is not supervised feature extraction, since the guidance comes from the input itself.

In neuroscience, many researchers have suggested that feedback (top-down, lateral or recurrent) connections, from other parts of the brain (context), have an important role in learning [see, e.g. Marr, 1982, Becker and Hinton, 1992, Parga and Rolls, 1998]. Recurrent feedback would correspond to temporal and lateral to spatial information.

An illustration of how the context can be used for learning is shown in Fig. 6. The crucial difference to the feature expansion is another denoising phase guided by the expectation-creating context. Then the sub-sequent PCA-stage finds directions where the driving input from the feature expansion has something in common with the context. Blocks with this structure can be assembled in a fully hierarchical DSS systems. One such is pictured in Fig. 7. The contextual input is denoted by colored lines: red for top-down, blue for recurrent and green for lateral.

Valpola [2004] has shown that spatial context alone is sufficient for developing invariant features (in SFA and related models temporal context has been used). We review the results here.

In the experiment, 9 laterally connected units were used on one layer. The contextual-guidance was realised by simply adding the context vector to the feature vector outputted from the feature expansion. 10% share of weights was assigned to the lateral contextual inputs.

The results are summarised in Fig. 8. In the top left, the whole image is shown. The data was

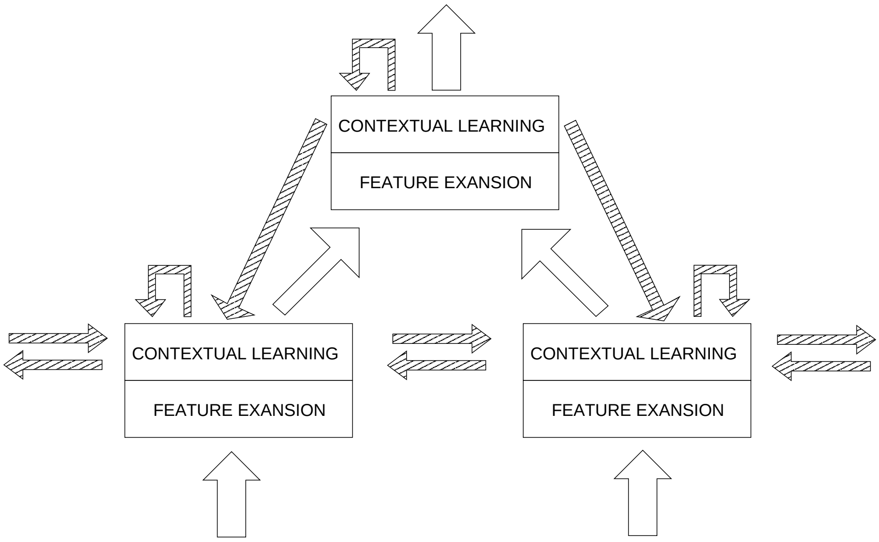


Figure 7: *Hierarchical feature extraction using DSS.*

7×7 patches sampled from it. Three such samples are shown in the rest of the top row. In the second row, there are activations of four out of 100 nonlinear features at the feature expansion level of one unit when scanning over the entire image. The third row shows the corresponding 7×7 receptive fields. The last row shows activations of a feature developed on the output layer under spatial contextual guidance. The strongest bottom-up connections were made to the four features shown on the previous rows. This feature is more invariant than any of the more elementary features.

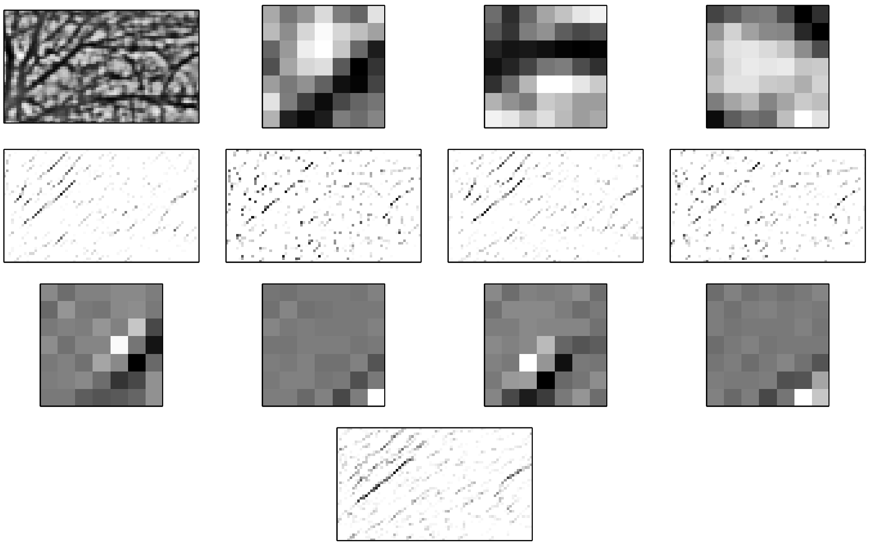


Figure 8: *Features extracted from a natural scene using context-guided DSS with spatial information.*

4.2 Role of attention in goal-directed learning

In the beginning we mentioned that one of the main purposes of the feature extraction is to facilitate the task at hand, such as object recognition or motor control. From psychophysical experiments it has become clear that attention plays a significant role in learning. Some researchers even claim that there is no learning without attention. This is controversial but it is nevertheless clear that attention is very important. In this section, we speculate how attentive mechanisms could be used in hierarchical DSS to guide learning. This hypothesis is more extensively discussed in [Valpola, 2004, 2005].

Contextual (predominantly top-down) biasing of local lateral competition has been proposed as a model of covert attention in humans [Duncan and Humphreys, 1989]. It seems that attention is realised by a competitive binding process that forms dynamically functional networks[Reynolds et al., 1999, Reynolds and Desimone, 1999]. This dynamical binding has been shown to gate the coherence between cortical areas, thereby affecting the associations learned between these areas [Miltner et al., 1999]. Moreover, attention is strongly influenced by goals and motivation.

We hypothesize that these featurs make attention a suitable mechanism for converting goals and motivation into teaching signals that guide the development of feature extraction. If attention can influence what type of information reaches a given cortical area, it can presumably also influence what kind of expectations are generated and thus control learning in context-guided DSS.

The role of attention in learning is illustrated in Fig. 9. The nodes represent different cortical areas, for instance sensory processing areas for different modalities or different parts of the visual field. The anatomical connections between the areas are represented by thin lines. The functional network dynamically created by attention is illustrated by the thick lines: these parts of the cortex are temporally functionally connected. Assume that the information passing in the primary cortical areas 1 and 2 are A and B, respectively. Primary cortical area 3 receives bottom-up inputs (not shown) which carry both types of information (A+B), so it could potentially try to extract both types of information. However, if attentional filtering routes either type of information to area 3 during learning, area 3 would be biased to extract that type of features in the future. Thus, if attention is focussed on A during learning, area 3 will learn to extract A-type information and vice versa. Specifically, our hypothesis is that attention can influence learning by modulating the expectations which in turn guide the development of feature extraction.

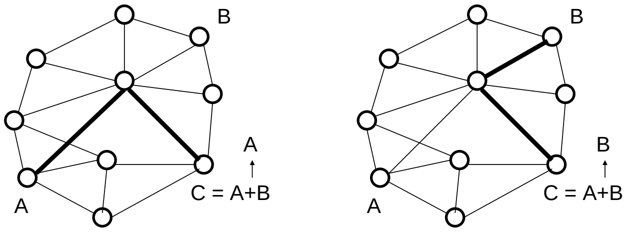


Figure 9: *Attention.*

Deco and Rolls [2004], Deco and Schürmann [2000] have concentrated on modelling the attention process without direct implication on how the representations would be learned. However, the structure of the model is compatible with Fig. 6 and our suggestion would thus integrate learning in the the attention process.

5 Conclusion

In this paper, we discussed denoising source separation (DSS) and especially its use for feature extraction. We showed that with simple combination of Hebbian and anti-Hebbian learning and suitable nonlinearity, powerful feature extractors can be realised. Additionally, we speculated how attention can be implemented in this framework. This would probably make it possible to implement even deep hierarchies of large-scale networks.

We have recently released a flexible MATLAB DSS package for source separation and feature extraction [DSS, 2004]. Most of the experiments in this paper can be implemented with it rather easily.

6 Acknowledgements

The research on nonlinear separation with MLP networks has been conducted in collaboration with Mariana Almeida (manuscript in preparation). JS is funded by the Academy of Finland (under the project New Information Processing Principles). HV is partly funded by the Academy of Finland (under the Research Centre for Computational Science and Engineering) and partly by European Commission (ADAPT project, IST-2001-37173).

References

L. B. Almeida and M. Faria. Separating a real-life nonlinear mixture of images. In Carlos G. Puntonet and Alberto Prieto, editors, *Proceedings of the Fifth International Conference on Independent Component Analysis and Signal Separation (ICA2004)*, volume 3195 of *Lecture Notes in Computer Science*, pages 729 – 736, Granada, Spain, 2004. Springer-Verlag, Berlin.

S. Becker and G. E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161 – 163, 1992.

T. Blaschke and L. Wiskott. Nonlinear blind source separation by integrating independent component analysis and slow feature analysis. In *Proc. Advances in Neural Information Processing Systems 17, NIPS'04*, 2005. in press.

G. Deco and E. T. Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision research*, 44:621 – 642, 2004.

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

DSS. The DSS MATLAB package. 2004. Available at `http://www.cis.hut.fi/projects/dss/`.

J. Duncan and G. Humphreys. Visual search and stimulus similarity. *Psychological Review*, 96:433 – 458, 1989.

M. H. Glantz. *Currents of Change: El Niño's Impact on Climate and Society.* Cambridge University Press, 1996.

P. Hoyer and A. Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191 – 210, 2000.

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

A. Hyvärinen, P. O. Hoyer, and J. Hurri. Extensions of ica as models of natural images and visual processing. In *Proceedings of the Fourth International Conference on Independent Component Analysis and Signal Separation (ICA2003)*, pages 963 – 974, Nara, Japan, 2003.

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. 2005 IEEE International Joint Conference on Neural Networks (IJCNN 2005)*, Montreal, Canada, 2005. To appear.

T. Kohonen, S. Kaski, and H. Lappalainen. Self-organized formation of various invariant-feature filters in the Adaptive-Subspace SOM. *Neural Computation*, 9(6):1321–1344, 1997.

D. Marr. *A Computational Investigation into the Human Representation and Processing of Visual Information.* W. H. Freeman, San Fransisco, 1982.

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*, 397(4, Feb.):434 – 436, 1999.

E. Oja. A simplified neuron model as a principal component analyzer. *J. of Mathematical Biology*, 15: 267–273, 1982.

E. Oja, H. Ogawa, and J. Wangviwattana. Learning in nonlinear constrained Hebbian networks. In T. Kohonen et al., editor, *Artificial Neural Networks, Proc. ICANN'91*, pages 385–390, Espoo, Finland, 1991. North-Holland, Amsterdam.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

N. Parga and E. Rolls. Transform invariant recognition by association in a recurrent network. *Neural computation*, 19:1507 – 1525, 1998.

J. H. Reynolds, L. Chelazzi, and R. Desimone. Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience*, 19:1736 – 1553, 1999.

J. H. Reynolds and R. Desimone. The role of neural mechanisms of attention in solving the binding problem. *Neuron*, 24:19 – 29, 1999.

J. Särelä and H. Valpola. Denoising source separation. *Journal of Machine Learning Research*, 6:233–272, 2005.

A. M. Thomson and A. P. Bannister. Interlaminar connections in the neocortex. *Cerebral Cortex*, 13:5 – 14, 2003.

H. Valpola. Behaviourally meaningful representations from normalisation and context-guided denoising. Technical report, Artificial Intelligence Laboratory, Department of Information Technology, University of Zurich, 2004. Available at Cogprints: `http://cogprints.ecs.soton.ac.uk/archive/00003633/`.

H. Valpola. Developments of representations, categories and concepts — a hypothesis. In *proceedings of CIRA 2005.* 2005. to appear.

H. Valpola and J. Särelä. Accurate, fast and stable denoising source separation algorithms. In *Proceedings of the Fifth International Conference on Independent Component Analysis and Signal Separation (ICA2004)*, pages 64 – 71, Granada, Spain, 2004.

L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14:715 – 770, 2002.