Online learning of feature detectors from natural images with the probabilistic WKL rule

Jasmin Léveillé
Center of Excellence for Learning in Education, Science, and Technology
Boston University
Boston, USA
jasminl@cns.bu.edu

Isao Hayashi[2] and Kunihiko Fukushima[1, 2]
[1] Fuzzy Logic Systems Institute, Iizuka, Fukuoka, Japan
[2] Faculty of Informatics, Kansai University
2-1-1, Ryozenji-cho, Takatsuki, Osaka, Japan
ihaya@cbii.kutc.kansai-u.ac.jp
fukushima@m.ieice.org

Abstract— **Recent advances in machine learning and computer vision have led to the development of several sophisticated learning schemes for object recognition by convolutional networks. One relatively simple learning rule, the Winner-Kill-Loser (WKL), was shown to be efficient at learning higher-order features in the neocognitron model when used in a written digit classification task. The WKL rule is one variant of incremental clustering procedures that adapt the number of cluster components to the input data. The WKL rule seeks to provide a complete, yet minimally redundant, covering of the input distribution. It is difficult to apply this approach directly to high-dimensional spaces since it leads to a dramatic explosion in the number of clustering components. In this work, a small generalization of the WKL rule is proposed to learn from high-dimensional data, and is shown to lead mostly to V1-like oriented cells when applied to natural images.**

Keywords—winner-kill-loser, competitive learning, incremental learning, clustering, natural image statistics

I. INTRODUCTION

Unsupervised learning techniques have been extensively used to learn features for convolutional neural networks in order accomplish difficult visual recognition tasks [1, 2]. Various unsupervised learning schemes have been proposed to learn good recognition features, including collecting a dictionary of features from an image dataset [3, 4], Hebbian-inspired learning [5], contrastive divergence [6], Independent Component Analysis [7, 8], and reconstruction error minimization [9].

Dictionary of features are built incrementally upon presentation of input patterns, by inserting new units that code for regions of the input space not covered by a current set of units. Upon insertion, a new unit's weights are set to encode the values of the input pattern currently applied and are fixed throughout the remainder of the learning phase. Although simple to implement, such incremental techniques can easily lead to prohibitively large dictionaries being learned on typical visual recognition tasks. Feature sets learned through dictionary building therefore run the risk of low coding efficiency since a large code is used to represent the data (cf. [10]).

Techniques based on Hebbian learning instead typically assume a fixed number of units whose weights are gradually learned upon repeated presentation of inputs with the product of pre- and post-synaptic activity. When used in conjunction with a suitable normalization operator and competitive interactions among feature units, the learned synaptic kernels correspond to either independent components [11] or principal components [12]. Hebbian learning can actually be related to reconstruction error minimization, in which synaptic weights are learned in order to minimize an objective function that measures the distortion between the input and the backprojected (i.e. through the synaptic matrix transpose) output [13]. One variation of Hebbian learning – the trace rule – uses low-pass filtered post-synaptic output with the hope of achieving spatial invariance through learning from continuously varying input [5, 14, 15]. Trace rule learning is sometimes equivalent to Slow Feature Analysis [16], in which the goal is to promote learning of features that capture slow variations in the input for purpose of invariance. Despite its simplicity and the possibly limiting influence of the linearity assumption of Hebbian learning, state-of-the-art results on standard image datasets have been achieved with this framework. Nevertheless, model selection remains an issue, such that the only viable strategy to obtain good results may be to rely on high throughput parameter search [5].

Hebbian-based, competitive learning has also been used in conjunction with a dictionary-building strategy in the Neocognitron model [17]. According to this scheme, existing units compete upon presentation of inputs, and the weights of the winning unit are modified according to a variant of Hebbian learning. New units are added whenever none of the existing units is sufficiently activated by a given training input. This method has shown good results in particular on digit recognition datasets.

Contrastive divergence has also been used to learn features for convolutional networks [6]. Learning with contrastive divergence roughly follows the gradient of the log-likelihood [18]. The results obtained with that approach on standard image datasets have been comparable to that obtained with other state-of-the art approaches. Here again, the number of units is fixed in advance and model selection is an issue. Learning with

This work was partially supported from International Program for Scholars from Overseas of Kansai University, Devision of International Affairs, and Kansai University by Strategic Project to Support the Formation of Research Bases at Private Universities: Matching Fund Subsidy from MEXT, 2008-2012.

Independent Component Analysis (ICA) is somewhat related to the maximum likelihood approach in that both can be traced back to optimizing a log-likelihood criterion [19], although ICA is intrinsically tied to the use of a sparsity criterion. Results obtained with ICA have been in line with that of other approaches on standard image datasets [7, 8].

Instead of performing gradient ascent on the log-likelihood function, another learning strategy is to perform gradient descent on the input reconstruction error in an autoencoder network [20]. Minimizing the reconstruction error can be shown to be equivalent to maximizing a lower bound on the mutual information between the input and output layers of the autoencoder [21]. Recent energy-based methods for learning convolutional autoencoders have yielded some of the best performing methods on standard object recognition benchmarks [1, 9].

Within the above approaches, the simplest learning rules are attractive for at least two reasons. First, simple learning mechanisms are more likely to be suitable for hardware acceleration, as shown by the fact that state-of-the-art hardware implementations of convolutional networks currently have to perform learning offline [22]. Second, learning mechanisms that rely purely on local computations and that do not require sophisticated numerical procedures are generally more adequate as explanations of learning in the biological brain [23].

In this article, we consider how one such simple learning rule, the Winner-Kill-Loser (WKL), can be used to learn oriented, V1-like receptive fields. The WKL rule was proposed in [17] for purpose of learning higher-order combinations of features in the neocognitron model for written digit classification [24]. The WKL rule can be described as follows. Let x be an input pattern, the activity of unit j is represented by a similarity function $s_j = f(w_j, x)$, where w_j is the unit's synaptic kernel. The similarity function f is typically implemented as a normalized dot product:

$$s_j = \frac{w_j \cdot x}{\|w_j\| \|x\|}, \tag{1}$$

where $\|\cdot\|$ is the L_2 norm. Let i be the index of the most active unit within a group of units tuned to different features. Given a certain activity threshold θ, and assuming $s_i > \theta$, the weight update for the winning unit is defined as:

$$\Delta w_i = \lambda x, \tag{2}$$

for a given learning rate λ. The threshold θ may be seen as defining a neighborhood in the input space, located around the center wi, for which patterns will strongly activate unit i. Reference [17] considered only the case of λ=1, although here we use λ=0.05 for all simulations. The weights for all inactive units (i.e. units whose activity is less than the threshold) are left unchanged. If no unit has an activity level beyond θ, a new node is created whose center is set to the input pattern x. Up to this point, the WKL is identical to incremental Winner-Take-All learning [25, 26]. However, the WKL includes an additional step whereby units whose activity is higher than the threshold but is less than that of the winner are removed from the network. WKL learning is thus a variant of incremental, competitive clustering techniques that include the leader algorithm [27] and adaptive resonance theory [28], and differs from these by the addition of a fast decremental step.

The purpose of the decremental step is to reduce redundancy in coding while allowing for a complete covering of the input space with fewer nodes than in traditional WTA learning, leading to a form of sparse coding of the input [29]. Simulations in the neocognitron have shown that the WKL compares favorably at least on a written digit recognition task [17].

One issue with incremental learning rules such as WKL is the potentially large number of nodes needed to cover the input space. With WKL learning, this problem appears in two different ways. First, the removal of existing nodes can be seen to lead to gaps in the covering, at least when using a spherical cluster neighborhood as in Eq.1. Second, the sheer size of the input space may simply be so high as to require a prohibitive number of nodes to cover its volume.

The first type of problem is not really of concern, as gaps between closely connected neighborhoods can be dealt with using simple strategies [30]. Furthermore, the general influence of these gaps can be expected to decrease as the dimensionality of the input space increases. That this is the case can be seen by considering the ratio of the volumes of the gaps to that of the nodes' neighborhood. For any dimension n, the spherical neighborhoods of three adjacent nodes form an equilateral triangle on a hyperplane. Let r be the radius of those spheres. Within that hyperplane, the three spheres define another one that covers approximately the projection, onto the hyperplane, of the gap at their intersections. The radius of that sphere is given by:

$$q = \frac{r}{\cos^{\pi}/_{6}} - r, \tag{3}$$

and its volume is:

$$V_n q^n, \tag{4}$$

where V_n is the volume of a hypersphere of radius 1 [31]. The ratio of the volumes of the gap to that of the nodes' neighborhoods is then:

$$V_n q^n / V_n r^n = \frac{\left(\frac{r}{\cos^{\pi}/_{6}} - r\right)^n}{r^n} = \left(\frac{1}{\cos^{\pi}/_{6}} - 1\right)^n. \tag{5}$$

It is easy to see that, as $n \to \infty$, this ratio converges to 0. Thus, as dimension increases, gaps between connected neighborhoods should have less impact on the covering. Fig.1b shows the result of Monte-Carlo simulations that confirm this intuition. Unlike the gaps between adjacent neighborhoods, increasing the input space dimensionality actually worsens the second type of covering problem (Fig.1a).

From a computational perspective, incremental competitive learning procedures thus seem inadequate to learn oriented receptive fields from the (high-dimensional) space of natural

images. Fig.2 shows an example of the kind of receptive fields learned by applying the WKL rule to natural images: few cells seem to display any form of orientation selectivity.

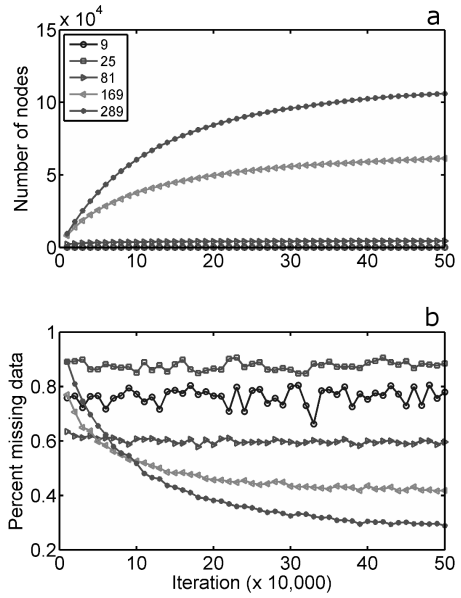

Figure 1. WKL learning and input space dimensionality. WKL simulations were run on randomly sampled natural image patches of increasing dimensionality (from 3x3 to 17x17). a) The number of nodes learned as a function of the number of learning iterations clearly increases for high-dimensional spaces but remains roughly constant at low dimensions. b) The resulting input space covering – as measured by the percent test data points not included in the covering – correspondingly improves for high-dimensional, but not low -dimension spaces.

Incremental learning procedures thus seem better suited to learn higher-order features formed by combinations of handcrafted oriented edge detectors like Gabor filters or Difference-of-Gaussians [4, 17].

II. PROBABILISTIC WKL

In this section we introduce a small generalization of the WKL learning rule that allows a useful form of incremental learning in high-dimensional space. In particular, the modification we propose emphasizes learning important statistical regularities in the data rather than attempting to cover the entire input space [32].

Our first modification to the original WKL rule consists in making the removal of a node inversely proportional to the number of times (n_i) it has won competition in the past. Let P_i denote the probability of removing node i given that its activity is above threshold but below that of the winning neuron. The model presented in [17] was restricted to the case where $P_i = 1$. What we propose is thus to use instead $P_i = 1/n_i$. The rationale for such a mechanism is that a node that has won many times in the past is likely to cover an important part of the input space and should thus be kept.

The second modification to the WKL rule consists in making new node insertion inversely proportional to the total number of nodes (N). Let P_n denote the probability of inserting a new node given that no existing node has an activity level beyond the threshold θ. The model presented in [17] was restricted to the case where $P_n = 1$, meaning that a new node is certain to be introduced. What we suggest instead is to use $P_n = 1/N^{\alpha}$, where alpha is a user-specified parameter. If no new node is inserted, the input pattern is nevertheless learned by the most active unit in the network, despite that its activity is less than θ. This second modification allows the network to maintain a relatively small number of nodes despite the potentially large dimension of the input space while maintaining the incremental nature of the original learning rule. Making node insertion inversely proportional to the number of existing nodes is analogous to the use of a Dirichlet prior in nonparametric Bayesian estimation [33], but without the complex sampling procedures required for statistical consistency [34]. Crucially, the use of a soft criterion in the WKL to determine new node insertion allows the number of nodes to adapt to the input distribution, unlike previous learning methods which considered a fixed number of units (e.g. [11, 23, 35, 36]).

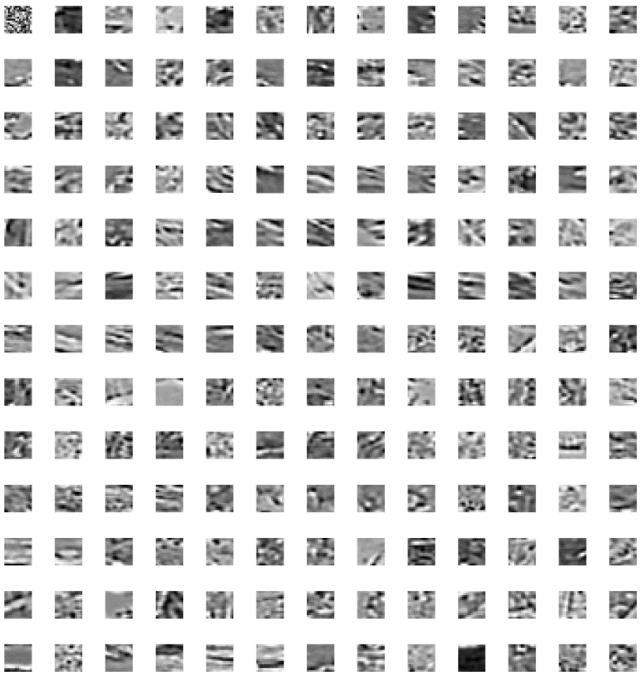

Figure 2. Example receptive fields learned with WKL learning applied to natural image data. Most cells do not develop a pattern of orientation selectivity.

III. LEARNING V1-LIKE RECEPTIVE FIELDS

In this section we demonstrate that the proposed generalization of the WKL rule is capable of learning oriented edge receptive fields from natural images. The training procedure is analogous to the ones used in typical V1 learning simulations. At each iteration, a 15-by-15 image patch is randomly gathered from a natural image and input to the network. For our simulations we use natural images gathered from a camera attached to the head of a cat as it wanders in a natural environment [37]. As in [38], raw pixel input is first pre-processed with a difference-of-Gaussians filter whose inner and outer spreads are given by 0.875 and 1.4, respectively. Cell activities are then computed

according to Eq.1 and learning proceeds as described above with an additional normalization of the input x in Eq.2 which further reduces the dimensionality of the input space, for a maximum number of 100,000 iterations. The threshold parameter is fixed at $\theta = 0.65$, and $\alpha = 1.4$.

Fig.3 shows the number of nodes learned as a function of the training iteration when using either the original WKL (thick black line) or its proposed generalization (pWKL; thin gray line). The number of nodes learned with the probabilistic WKL is clearly inferior – by a few orders of magnitude – to the number learned with the original WKL. The network also appears to be approaching stability, although simulations run with an even higher number of training patterns should be conducted to confirm whether stability is achieved.

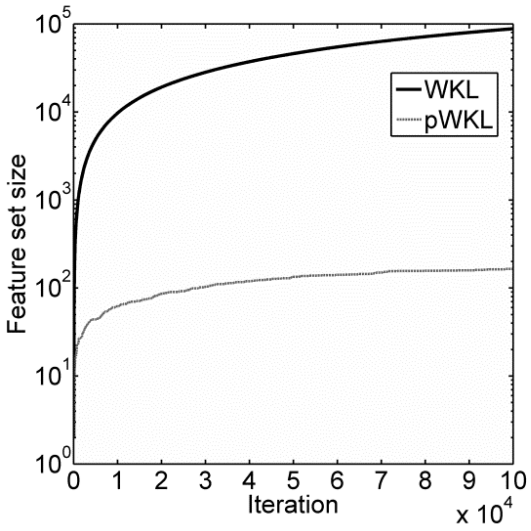

Figure 3. Comparison of the number of nodes learned with the original WKL and the probabilistic WKL (pWKL) during training. The number of nodes learned with pWKL is clearly less than with the original learning rule.

Fig.4 shows the first 169 receptive fields (out of a total of 181) learned with the probabilistic WKL. In comparison to the patterns learned with the original WKL (Fig.2), the probabilistic WKL appears to capture important regularities in the natural image input space.

In order to quantify how well the learned receptive fields approximate the near uniform distribution of edges in natural scenes, Gabor patches were fit via least-squares, and the resulting phase/frequency estimates visually inspected for correctness. In 7% of cases the fitting procedure failed to converge despite that a clear orientated receptive field had been learned. In 4% of cases, the receptive fields obtained lacked a clear orientation, and instead resembled the kind of center-surround receptive fields found in undirected V1 cells. Finally, in 14% of cases the learning algorithm failed to yield a receptive field with a clearly identifiable structure.

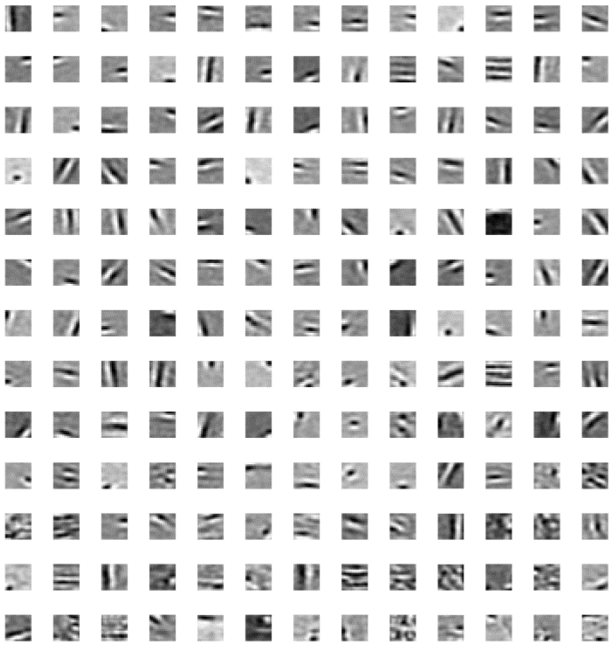

Figure 4. V1-like receptive fields learned with pWKL. Orientation selectivity can easily be observed in a majority of units.

The estimates obtained for all remaining receptive fields are plotted on the log-polar plane in Fig.5.

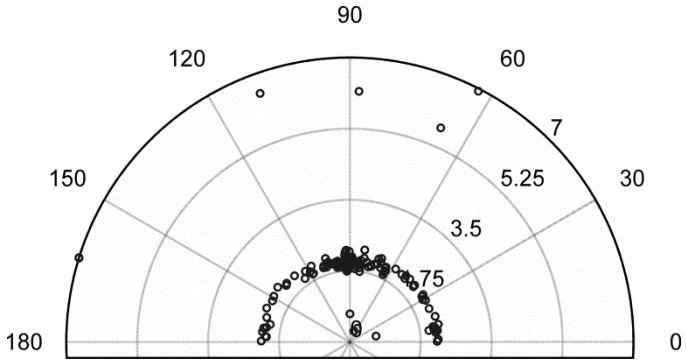

Figure 5. Log-polar representation of the distribution of phase and frequency of filters learned with pWKL. Phase is indicated by the angle on the polar plot. Log-frequency is indicated by the radial circles. All phases are adequately represented, with a particular emphasis on vertical orientations.

As Fig.5 shows, the learning algorithm is able to successfully learn edge filters in all orientations. The learning rule is not as successful at spanning the space of frequencies. More investigations are needed to determine whether this limitation should be ascribed to the learning rule itself or to the pre-processing used here (in particular, to the spatially frequencies of the difference-of-Gaussians filter).

IV. Discussion

Our primary objective in this article is to see whether a simple incremental learning rule such as WKL can be used to learn statistical regularities in the high-dimensional space induced by natural images. Upon presentation of natural images, in its

original version, the WKL rule leads to an explosion in the number of learned filters (Fig.3). In addition, most learned filters do not capture essential regularities in the data (Fig.2). On the other hand, the WKL rule learns these regularities when generalized so as to make a given node's removal probability inversely proportional to the number of times it has won competition, and so that new node insertion is inversely proportional to the total number of nodes. The WKL rule learns not only strongly oriented receptive fields, but also, to a lesser extent, undirected receptive fields. The latter resemble the zero-phase (ZCA) filters of [35], and their emergence among a wider group of oriented cells is consistent with the fact that non-oriented, black-white cells are also present in cortical area V1 [39]. Fig.4 reveals that learned undirected cells were of the off-center on-surround type exclusively. Future work is needed to understand why this type of cells was learned over on-center off-surround cells.

Both proposed generalizations require minimal changes to the original WKL. In particular, the quantity P_i requires only evaluating local computations. The quantity P_n requires knowing the total number of nodes N which, despite being a global quantity, remains simple to compute. Although this question is beyond the scope of this article, it is possible that such a quantity would be computed implicitly in the brain by considering that the number of neurons in a given cortical volume remains roughly constant.

In this work, the probabilistic terms P_i and P_n were kept as simple as possible. It remains a possibility, however, that the postulated forms limit the range of learned features. For example, using a heavy-tailed function for P_n might lead to more useful features being learned. Conversely, stable learning may be impaired due to the fact that nodes are initially prone to removal due to the high value of P_i.

Although the space of orientations is appropriately covered by the learning rule (Fig.5), variations in frequency do not seem to be well handled by the learning rule. Such a tight clustering of learned frequencies has already been observed when using either independent component analysis (ICA) or sparse coding techniques [40]. It is not clear at present why such a tight clustering would occur.

V. CONCLUSION

In this article, we generalize the WKL rule to learn oriented receptive fields from natural image data. Our generalization retains some of the essential benefits of the original WKL – namely its simplicity and biological plausibility – while being able to deal with a high-dimensional input space. The main drawback of the proposed method is that it does not lead to a dense covering of frequency space.

REFERENCES

[1] K. Jarrett, K. Kavukcuoglu, M.-A. Ranzato and Y. LeCun, "What is the best multi-stage architecture for object recognition?, " Proc. ICCV, pp.2146-2153, 2009.

[2] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition, " Proc. of the IEEE, pp. 2278-2324, 1998.

[3] J. Mutch and D. G. Lowe, "Multiclass object recognition with sparse, localized features," Proc. CVPR, pp. 11-18, 2006.

[4] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber and T. Poggio, "Robust Object Recognition with Cortex-like Mechanisms," IEEE Transactions on Pattern Analysis and Machine Intelligence, 29, pp. 411-426, 2007.

[5] N. Pinto, D. Doukhan, J. J. DiCarlo and D. D. Cox, "A high-throughput screening approach to discovering good forms of biologically inspired visual representations," PLOS Computational Biology, 5, e1000579, 2009.

[6] H. Lee, R. Grosse, R. Ranganath and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," Proc. ICML, pp. 609-616, 2009.

[7] H. Lee, E. Chaitanya, and A. Y. Ng, "Sparse deep belief network for visual area V2," NIPS, 2007.

[8] Q. V. Le, J. Ngiam, Z. Chen, D. Chia, P. Koh and A. Y. Ng, "Tiled convolutional neural networks," NIPS, 2010.

[9] M.-A. Ranzato, F.-J. Huang, Y.-L. Boureau and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," Proc. CVPR, 2007.

[10] P. D. Grünwald, "The Minimum Description Length Principle," MIT Press, 2007.

[11] N. Zhang and J. Weng, "Sparse representation from a winner-take-all neural network," Proc. IJCNN 2004, pp. 2209-2214, 2004.

[12] E. Oja, "Simplified neuron model as a principal component analyzer," Journal of Mathematical Biology, 15, pp. 267-273, 1982.

[13] J. L. Jr Wyatt, and I. M. Elfadel, "Time-domain solutions of Oja's equations," Neural Computation, 7, pp. 915-922, 1995.

[14] P. F□ldiák, "Learning invariance from transformation sequences," Neural Computation, 3, pp. 194-200, 1991.

[15] E. T. Rolls and T. Milward, "Model of Invariant Object Recognition in the Visual System: Learning Rules, Activation Functions, Lateral Inhibition, and Information-Based Performance Measures," Neural Computation, 12, pp. 2547-2572, 2000.

[16] H. Sprekeler, C. Michaelis and L. Wiskott, "Slowness: An Objective for Spike-Timing–Dependent Plasticity?," PLoS Comput Biol 3, 2007.

[17] K. Fukushima, "Neocognitron trained with winner-kill-loser rule," Neural Networks, 23, pp. 926-938, 2010.

[18] G. Hinton, "Training product of experts by minimizing contrastive divergence," Neural Computation, 14, pp. 1771-1800, 2002.

[19] A. Hyvärinen, J. Hurri and P. O. Hoyer, "Natural Image Statistics – A probabilistic approach to early computational vision," Springer-Verlag, 2009.

[20] G. W. Cottrell, P. Munro, and D. Zipser, "Learning internal representations from gray-scale images: An example of extensional programming," Proc. Cognitive Science Society, 1987.

[21] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," Journal of Machine Learning Research, 11, pp. 3371-3408, 2010.

[22] C. Farabet, Y. LeCun, K. Kavukcuoglu, E. Culurciello, B. Martini, P. Akselrod, and S. Talay, "Large-scale FPGA-based convolutional networks," In Bekkerman, Ron and Bilenko, Mikhail and Langford, John (Eds), Scaling up Machine Learning: Parallel and Distributed Approaches, Cambridge University Press, 2011.

[23] L. N. Cooper, N. Intrator, B. S. Blais and H. Z. Shouval, "Theory of cortical plasticity," Singapore: World Press Scientific, 2004.

[24] K. -Fukushima and S. Miyake, "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position," Pattern Recognition, 15, pp. 455–469, 1982.

[25] S. Grossberg, "Contour enhancement, short-term memory, and constancies in reverberating neural networks," Studies in Applied Mathematics, 52, 1973.

[26] T. Kohonen, "Self-organized formation of topologically correct feature maps," Biological Cybernetics, 43, pp. 59-69, 1982.

[27] J. A. Hartigan, "Clustering algorithms," New York: John Wiley & Sons Inc, 1975.

[28] S. Grossberg, "Competitive learning: From interactive activation to adaptive resonance," Cognitive Science, 11, pp. 23-63, 1987.

[29] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature, 381, pp. 607-609, 1996.

[30] K. Fukushima, I. Hayashi and J. Léveillé, "Neocognitron trained by winner-kill-loser with triple threshold," ICONIP, 2011.

[31] J.H. Conway and N. J. A. Sloane, "Sphere packing, lattices and groups," New York: Springer-Verlag, 1988.

[32] G. Hinton, "To recognize shapes, first learn to generate images," Progress in Brain Research, 165, pp. 535-547, 2007.

[33] Y. W. Teh, "Dirichlet processes," In Encyclopedia of Machine Learning. Springer, 2010.

[34] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," Journal of Computational and Graphical Statistics, 9, pp. 249-265, 2000.

[35] A. J. Bell and T. J. Sejnowski, "The independent components of natural scenes are edge filters," Vision Research, 23, pp. 3327-3338, 1997.

[36] R. Mikkulainen, J. A. Bednar, Y. Choe and J. Sirosh, "Computational maps in the visual cortex," Springer, 2005.

[37] B. Betsch, W. Einhäuser, K. Körding and P. König, "The world from a cat's perspective – statistics of natural videos," Biological Cybernetics, 90, pp. 41-50, 2004.

[38] T. Masquelier, T. Serre, S. J. Thorpe and T. Poggio, "Learning complex cell invariance from natural video: a plausibility proof," CBCL Paper. Massachusetts Institute of Technology; Cambridge, MA, 2007.

[39] M. S. Livingstone and D. H. Hubel, "Anatomy and physiology of a color system in the primate visual cortex," Journal of Neuroscience, 4, pp. 309-356, 1984.

[40] Y. Karklin and M. S. Lewicki, "Is early vision optimized for extracting higher-order dependencies?," NIPS, 2005.