Computational Models of Machine Vision
Goal, Role and Success

Tayyaba Azim and Mahesan Niranjan
Communications, Signals, Processing and Control (CSPC) Group,
School of Electronics and Computer Science, University of Southampton, Southampton, U.K.

Keywords: Object Detection, Feature Learning, Deep Models, Support Vector Machines, Fisher Kernel.

Abstract: This paper surveys the learning algorithms of visual features representation and the computational modelling approaches proposed with the aim of developing better artificial object recognition systems. It turns out that most of the learning theories and schemas have been developed either in the spirit of understanding biological facts of vision or designing machines that provide better or competitive perception power than humans. In this study, we discuss and analyse the impact of notable statistical approaches that map the cognitive neural activity at macro level formally, as well as those that work independently without any biological inspiration towards the goal of developing better classifiers. With the ultimate objective of classification in hand, the dimensions of research in computer vision and AI in general, have expanded so much so that it has become important to understand if our goals and diagnostics of the visual input learning are correct or not. We first highlight the mainstream approaches that have been proposed to solve the classification task ever since the advent of the field, and then suggest some criterion of success that can guide the direction of the future research.

1 INTRODUCTION

Artificial object recognition has for long remained an important problem in computer vision because of its wide applications and the persistent gap in the performance between the humans and artificial scene recognition systems. The capability of human visual system supersedes machine vision in many respects. For example the *very large* number of object categories that humans learn and recognize *accurately* despite the variability in their position, size, viewpoint, illumination, clutter and distractions, motivates us to understand the functionality of human perception and emulate it in artificial systems. However, it is not easy to accomplish this task because of the limited neurophysiological findings on how brain learns and organizes the visual input. Fortunately, the advances in both vision algorithms and hardware have made practical visual object recognition within reach, as can be seen in systems deployed on airports and highways for security and risk assessments. However, a comprehensive solution to this problem where machines are as good as humans in terms of accuracy and speed of recognition, still evades the reach of even the best researchers with only partial solutions and limited success in constrained environment being the state of the art (Aggarwal et al., 1996).

One of the straightforward ways of refuting a theory of how humans learn is to show that the machine predictions do not match the given data. This is why many recognition algorithms are assessed based on their *prediction accuracy* and *precision* on benchmark data sets. Another approach to deal with this issue is to check whether the *internal representations* of visual input match that of the brain. This line of thought has lead to various biologically inspired feature detectors, graphical models and learning algorithms that learn features resembling physiological signals of the brain. In this paper, we first of all provide a survey of the mainstream approaches of object recognition algorithms and then suggest some cues that might help in guiding the progress of object recognition systems in future.

2 EARLIER COMPUTATIONAL MODELS OF OBJECT RECOGNITION-ERA 1930-2005

Some of the early ideas on computational modelling of visual object recognition can be traced back to Gestalt's work in the 1930s (Wertheimer, 1938), (Fulcher, 2003) that describes how visual objects can

Azim T..
Computational Models of Object Recognition - Goal, Role and Success.
DOI: 10.5220/0004737601790186
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 179-186
ISBN: 978-989-758-003-1

be separated from each other and from the background. The Gestalt psychologists maintained that humans constantly search for a 'good fit' between the visual image and the stored memories of visual objects that are naturally organized in the brain as patterns based on their continuity, similarity, closure, proximity and symmetry. These defined principles of perception that assist grouping of stimuli and were minimally effected by an individual's past experience are known as the *Laws of Pragnanz*. Gestalt theory laid much of the groundwork for the study of object recognition, however it was criticized heavily because of being more *descriptive* than *explanatory* enough to clarify the functioning of human vision. More tangible work that draws ideas from the study of neural processing in computational form, comes from Hubel and Wiesel (Hubel and Wiesel, 1962), (Hubel and Wiesel, 1965), who by observing the cat's visual cortex introduced the concept of hierarchical visual information processing in the receptive fields that are regions where the action of light causes reflex in the neurons. According to their findings, the receptive fields of cells at one level of the visual system are formed from input by cells at a lower level of the visual system. In this way, small, *simple* receptive fields could be combined to form large, *complex* receptive fields thus accounting for a progressive increase in the complexity of physiological receptive fields of cells. It was this discovery of receptive fields and *feed forward architecture* that later on lead to the development of many different hierarchical models of machine vision (Fukushima, 1988), (Wallis and Rolls, 1996), (Riesenhuber and Poggio, 1999), (Deco and Rolls, 2006).

One of the most influential work on understanding visual scene analysis after Hubel and Wiesel is that of David Marr, who proposed hierarchical modelling of the visual system from simple to complex at three independent levels of abstraction: *computational, algorithmic and implementational* levels. In computer analogy, these can be roughly understood as task, software and hardware levels. According to Marr, separating the three levels allow those interested in cognition to focus on the level they are most interested in, while simultaneously allowing those not specially interested in cognition (like computer scientists) to provide valuable insight from their specific point of view. The *tri-level hypothesis* is not without any objections, but it remained a valuable tool to aid in the study of cognitive science (and cognition in general). Apart from the *tri-level hypothesis*, Marr also proposed an intermediate stage of information representation - the 2-1/2D sketch - between the 2D image on the retina and a 3D description of the world in cortex (Marr and Poggio, 1979), (Grimson, 1981). The idea of a primal sketch is similar to a pencil drawing by an artist in which different areas of a scene are shaded to give depth to it. This *bottom up hierarchical processing* insight although seminal, has now been modified by the recent research (Serre et al., 2007b), (Rolls et al., 2009). However, it highly influenced the state of the art object recognition systems circa 1965-1980, giving birth to *object centered/shape based models* that focused on finding the correct representation for visual primitives, and represented objects hierarchically in terms of their structural properties. Marr and Nisihara's idea of part based structural representation was based on hierarchically stored three dimensional volumes of generalized cones (or cylinders) and their spatial relationship to one another (Marr and Nishihara, 1978). This approach of using geometric primitives was an attempt to reconstruct the shape of objects, in a similar vein to how some other inspired approaches in parallel line of work (Nevatia and Binford, 1973) were trying to reconstruct the scene, however, they did not provide any empirical support for the proposed model. Compared to this initially proposed model by Marr, the most well received structural descriptive model was the *recognition by components (RBC)* model by Biederman (Biederman, 1987) who refined Marr's model of object recognition in important ways and provided empirical support for the proposed theory : First and foremost improvement was the psychophysical support of the RBC model (Biederman, 1986), (Biederman and Cooper, 1991). Another defining factor of the recognition by components (RBC) theory was its ability to recognize the objects regardless of the viewing angle, known as *viewpoint invariance*. Although this model made sensible assumptions of how human visual system may parse a scene, it was not without caveats in practice. The main disadvantages of such shape-based methods are: the dependency on reliable extraction of geometric primitives (lines, circles, etc.), the ambiguity in interpretation of the detected primitives (presence of primitives that are not modelled), the restricted modelling capability owned by a class of objects which are composed of few easily detectable elements, and the need to create the models manually (Matas and Obdrzalek, 2004).

In contrast to the view point independent method proposed by Biederman, the decade of 1990s saw the evolution of *appearance/view based models* in which the objects are represented with respect to their viewpoint, thus entailing multiple representations that place higher demands on memory capacity; however it does potentially reduce the degree of computation necessary for deriving higher-level object represen-

tations in object centered models. Based on the derived features, these methods can be sub-divided into two main classes, i.e., local and global approaches. A local approach grabs a feature from a small region of an image (object) which is ideally a distinctive property of the object's view/projection to the camera. Examples of local features of an object are the color, mean gradient or mean gray value of pixels from small region. In contrast, the global approaches grab features that cover the information content of the whole image. This varies from simple statistical measures (e.g., mean values or histograms of features) to more sophisticated dimensionality reduction techniques, i.e., subspace methods, such as principle component analysis (PCA), independent component analysis (ICA), or non negative matrix factorization (NMF). Some of the popular methods that come into this category of models were proposed by (Turk and Pentland, 1991), (Linsker, 1992), (Lades et al., 1993), (Ojala et al., 1994), (Murase and Nayar, 1995), (Bell and Sejnowski, 1997), (Lowe, 1999). The question of whether the human visual system uses a *view based* or an *object centered* representation has been a subject of much controversy (for reviews see references (Logothetis and Sheinberg, 1996) and (Tarr and Blthoff, 1998)). We will just mention here the fact that the psychophysical and physiological data from humans and monkeys actually supports a view based approach. View/appearance based approach is attractive since it does not require image features or geometric primitives to be detected and matched. But their limitations, i.e. the necessity of dense sampling of training views and the low robustness to occlusion and cluttered background, make them suitable mainly for certain applications with limited or controlled variations in the image formation conditions, e.g. for industrial inspection (Matas and Obdrzalek, 2004).

In order to address the issues faced by object centered and appearance based models, *feature based methods* were proposed next, in which the objects are represented by a set of view independent local features which are automatically computed from the training images and stored as a database for probing the class of the test images later. Putting local features into correspondence is an approach that is robust to object occlusion and cluttered background in principle. Thus, when part of an object is occluded by other objects in the scene, only features of that part are missed and as long as there are enough features detected in the unoccluded part, the object can be recognized. Examples of such features that have been widely used for object recognition are scale invariant feature transform (SIFT) (Lowe, 2004), histogram of gradients (HoG) (Dalal and Triggs, 2005), haar wavelet feature set (Viola and Jones, 2001), etc. Such local patch based methods hold biological plausibility and tend to show benefits over global approaches when supported by mathematical models and neural network frameworks in object categorization (Leibe and Schiele, 2003).

One of the most popular ways of transforming a set of low level features extracted from an image into a high level image representation is the *bag of visual words (BoW)* inspired by the traditional bag of words technique for text analysis. The BoW algorithm constructs a codebook analogous to a dictionary from the collection of orderless patch based features, where each codeword in the codebook is a representative of several similar patches attained through the clustering process; consequently the test image can be represented by the histogram of the codewords. Several state-of-the-art visual object recognition systems (Csurka et al., 2004), (Zhang et al., 2007), (Li et al., 2008), (Wu and Rehg, 2011) fit into this general framework of codebook based object recognition models. After the image features are represented in the codebook of bag of words model, learning and recognition can be done in a *generative* or *discriminative* way. One of the greatest challenges in building up a codebook based model is the computation time required for clustering million of feature data points. (Ramanan and Niranjan, 2010) proposed a solution to this problem by presenting a sequential one-pass algorithm that creates the codebook in a drastically reduced time.

From a computational scientist's point of view, it goes without saying that this continuous research effort of developing bio-inspired architecture, learning algorithms and features was only taking place because the machines had not achieved human compatible speed and accuracy of detecting scenes. In this respect, the first researcher to quantify the timing of the visual scene understanding in humans was Simon Thorpe (Thorpe et al., 1996), who explained through event related potentials (ERP) analysis, the amount of time it takes to categorize the visual scene in cortex, which is 150ms. Progress towards understanding object recognition was driven by exploring and linking phenomenon at different levels of abstraction. At one end, where hierarchical generative models and learning algorithms inspired from the cortex were being improved, statistical methods independent of the biology of the visual system were also being developed in parallel. One popular paradigm which gathered a lot of attention since mid 1990s was the Vapnik theory of support vector machines (SVM) which showed impressive classification performance on many bench-

mark data sets. SVMs utilize a principle called *kernel trick* that computes dot products in high dimensional feature spaces using simple functions called *kernels* defined on pairs of input patterns. This trick enables us to get a linearly separable hyperplane for the data which is otherwise nonlinearly separable in the input space. Not only did the SVM classifier work successfully with the state of the art BoW feature space (derived from the BoW model discussed just above), but also with Fisher kernels (Jaakkola and Haussler, 1998) that combined the benefits of generative and discriminative approaches to pattern classification by deriving a kernel from a generative model of the data. Kernel classifiers like SVM proved their significance in various applications but they require a large amount of labelled training data as well as aprior definition of a suitable similarity metric/feature space in which naive similarity metrics suffice the classification to perform well. This requirement invites criticism by the researchers who are of the view that arranging a large amount of labelled data for many objects is expensive/impractical.

Although most of the proposed object recognition systems are inspired from the hierarchical nature of the primate cortex, it is worth mentioning that the *neural connectivity* and *learning algorithms* of these models have evolved with time. Earlier, most of the computational efforts were focussed on feed forward processing of information but since these feed forward connections just constitute a small fraction of the total connectivity in cortex, researchers shifted their attention towards the development of systems that made use of the back projection feedback too (Rumelhart et al., 1986). Feedback using back projection provides the opportunity of using previous knowledge, memory and task dependent expectations in a system (Kreiman, 2008), (Karklin and Lewicki, 2005). This change in neural connectivity revolutionized the learning algorithms used in undirected graphical models (Rumelhart et al., 1986), directed graphical models (Hinton et al., 1995) and non graphical models (Rumelhart et al., 1986). Although these theories failed to answer the scientific question of how the brain learns visual features, they produced two neat tricks: one for learning directed graphical models (Thulasiraman and Swamy, 1992) and the other one for undirected models (Karklin and Lewicki, 2009), (Hinton et al., 2006). Another influential fact that was established was that individual neurons are not sufficient for discriminating between objects; rather *population* of neurons should be analysed - a neuronal behavior also pointed out by the Wilson-Cowan model (Wilson and Cowan, 1972) in early 1970s and later addressed in many computational neuroscience problems (Sejnowsky, 1976), (Amit and Brunel, 1997), (Brunel, 2000), (Hertz et al., 2004).

3 RECENT COMPUTATIONAL MODELS OF IMAGE UNDERSTANDING-ERA 2006-PRESENT

Much of the progress experienced in the last decade has produced an overwhelming body of object recognition results without explaining anything significant about the perception and vision phenomenon in human visual system. The success of these artificial systems is determined by the overall recognition *accuracy* and the *time* they take to categorize these images. In order to cater the speed of recognition, it is worth mentioning the spectacular success of Poggio and Serre (Serre et al., 2007a) for developing an *immediate recognition system*, which is the fastest known form of computer object recognition against humans. In this system, the parallel processing paradigm was implemented rather than the conventional serial processing machine learning. When analysed with animal presence/absence test with humans, the computer did as well as the humans, and thus better than the best machine vision programs available so far. Immediate object recognition laid a new foundation of overall visual recognition and extending this theory to solve harder perception problem requires recruiting higher levels of brain function which would take more time and computational complexity for implementation. This extension has already began to spread in the Neuroscience community; an example being Stan Bileschi who applied this model to scene recognition (Bileschi, 2006), which is a derivation of higher order judgements, like it is a farm, a barn, a forest, etc.

As far as the goal of gaining better accuracy is concerned, *deep learning and representation* has been the subject of much recent research ever since the proposed breakthrough in feature learning by Hinton in 2006 (Hinton et al., 2006). The central idea of his greedy layer wise pre-training procedure is based on training each layer of the graphical model independently in an unsupervised way and then taking the features learnt at the previous layer as input to the next level. The features learnt by the deep model can either be used as an input to a standard supervised machine learning predictor such as support vector machines or as an initialization for a deep supervised neural networks like multi-layer perceptron (MLP). This idea of greedy layer wise unsupervised training was followed up quickly by the rest (Hinton and

Nair, 2009), (Taylor and Hinton, 2009), (Krizhevsky et al., 2012) as deep architectures showed potential of progressively learning more abstract features at higher levels of representation, yielding better classification error (Larochelle and Bengio, 2008), (Erhan et al., 2010) quality of samples generated by the probability distributions (Hinton and Salakhutdinov, 2009) and the invariance of properties learnt by the classifier . The recent work of (Krizhevsky et al., 2012) also shows that with proper initialization of parameters and choice of non linearity, it is not necessary to do unsupervised pretraining of the model as required by other deep networks.This finding reinforces the hypothesis that the unsupervised pretraining acts as a prior that brings little/no improvement over pure supervised learning from scratch when training data is large. The deep learning algorithms first proved their dominance over the MNIST digits data set by breaking the SVMs classification supremacy, and then moved on to object recognition in natural images. The latest breakthrough has been achieved on the Image net data set, bringing the error rate of the state of the art algorithms from 26.1% to 15.3% (Krizhevsky et al., 2012) on 10K classes of objects. While deep learning algorithms are making influential progress, another impressive approach making its mark in parallel is that of Fisher kernels with SVMs (Jaakkola and Haussler, 1998). The Fisher kernels made a successful come back by first showing its classification advantage over the state of the art bag of words approach (Perronnin and Dance, 2007), (Perronnin et al., 2010), and then showing its successful use with large scale data sets like PASCAL VOC 2007 (Csurka and Perronnin, 2011), CALTECH-256 (Sanchez and Perronnin, 2011) and ImageNet-10K (Sanchez et al., 2013). Currently, the second best performance after deep convolution network (Krizhevsky et al., 2012) achieved on the Image Net 10K classification task is shown by the Fisher kernels (Sanchez et al., 2013) derived from a gaussian mixture model designed for SIFT, local binary pattern (LBP) and GIST descriptors of the data.

4 SOME GUIDING PRINCIPLES FOR THE PROGRESS OF OBJECT RECOGNITION

In this section, we point out some misleading practices by the research community in the field of computer vision and introduce some novel aspects of research that have either been ignored completely or given less attention so far. We maintain that taking care of these aspects might improve the progress of artificial object recognition systems in the future.

- **Evaluation of the Benchmark Data Sets**
  In order to evaluate the strength of the learning algorithms and performance of classifiers, the experiments are usually conducted on standard benchmark data sets for comparison. Pinto (Pinto et al., 2008) argued that publicly available data sets such as Caltech-101 and PASCAL VOC image sets lack in several aspects that can actually mislead the progress in the long-term interest of being able to achieve near human levels of recognition. To prove this claim, he carried out the experiments on a V1 like model which was based on the known properties of simple cells of primate visual cortex. The model was a population of locally normalized, thresholded Gabor functions spanning a range of orientations and frequencies. This model contained no explicit mechanism to tolerate variation in object position, size/pose and shape. A standard one-versus-all approach was used to generate the multi-class SVM classifier from the training images. It was found that this V1 like model performed remarkably well on the Caltech-101 data set but when tested on a carefully controlled object recognition task that just consisted of two classes, the problem proved substantially harder for the V1 like model, exactly as one would expect for an incomplete model of object recognition. This proved that the V1 like model performed well previously not because of it being a good model of object recognition but because the natural image sets were inadequate. Ponce et al. (Ponce et al., 2006) also pointed out some of the issues present in the current standard data sets (i.e. UIUC, Caltech-4 and Caltech-101) used for judging the performance of developed object recognition systems. The most commonly observed problems in all these data sets were the limited range of variability in viewpoint, orientation of different instances in each category, no occlusion and background clutter. We have not seen any work that objects these claims about the inadequacy of these standard data sets or provides a counter solution to this problem. We suggest that there should be a formal mechanism of assigning a *challenging score* to each of the benchmark data sets in practice; based on this measure, the ones that are too simple should be discarded for experimentation in the future. Such an initiative is important to provide a uniform test bed to all the competing algorithms on fair basis of evaluation defined explicitly through the challenging score.

- **Impact of Learning Algorithms, Features and Amount of Training Data**
  The object/scene classification approaches often focus on one of the three aspects of the recognition problem: the amount of training data, the efficiency of learning algorithm and the quality of feature representations. It is important to know which of these factors are responsible for humans superior classification performance. The answer to this question was investigated by (Parikh and Zitnick, 2010), who compared the human and machine responses on similar problems to evaluate which of the three factors: learning algorithm, amount of training data and features, are responsible for better performance. They found no evidence that human pattern matching algorithms are better than standard machine learning algorithms. Also humans do not take advantage of increased amount of data, thus the main factor impacting the accuracies is the choice of features. we maintain that these observations should be further investigated and not ignored in order to focus the efforts in the right direction.
- **Integration between Physiological Recordings and Empirical Results of Object Recognition**
  Learning systems inspired by the biology and evaluated by their classification performance have become much more sophisticated in the last few decades. However, there is a need to directly verify the empirical results of machine recognition algorithms with the physiological recordings. Physiological data may offer an avenue for recognizing aspects of recognition that may be less obvious for humans but more suitable for computers. Such recognized cues could be integrated within a machine's control architecture to make it more capable of responding to visual signals in real time.
- **Addition of Time Dynamics**
  Most of the well known computational models reviewed here do not take into explicit account the fact that the retinal input usually has a time component associated to it. It is important to consider the time dynamics of the neural circuit as the objects in our surroundings move and the eyes show movement as well. Thus, measured neuronal responses are functions of time and even for an image presented in a flash, different types of information is carried out over time (Perrett and Oram, 1993), (Sugase et al., 1999) or in the time structure of the neuronal response. Incorporating the time dimension in neuronal models of recognition is a challenge that began in the last decade and is now actively being persued (Reichert et al., 2011b), (Reichert et al., 2011a). One of the interesting work in this regard is of (Nishimoto et al., 2011) who experimented on reconstructing the visual brain activity elicited by natural scene movies in humans. The time dynamics of the system is captured through a motion-energy model that describes how spatial and temporal information are represented in voxels throughout the visual cortex and then uses a Bayesian approach to combine estimated encoding models with a sampled natural movie prior for movie reconstruction. Under the Bayesian framework, the probability that the image *s* evoked response *r*, is given by the posterior distribution, $p(s|r)$ as follows: $p(s|r) \propto p(s)\prod_i p_i(r_i|s)$. To produce a reconstruction, $p(s|r)$ is evaluated for a large number of images. The image with the highest posterior probability, $p(s|r)$ is selected as the reconstruction, a method commonly known as *maximum a posteriori(MAP)* estimate. Much of the excitement surrounding this work is motivated by the ultimate objective of directly picturing subjective mental phenomenon such as visual imagery (Thirion et al., 2006) or dreams. We argue that *time* is an interesting dimension of the data, which if added to the existing models, can assist in making interesting discoveries about the human vision that could be deployed in the artificial systems.

5 CONCLUSIONS

The huge successes of the *Annual Meeting on Object Perception, Attention, and Memory* (in its 21st year), *Proceedings of the Neural Information Processing Systems* (in its 26th year) and the *Annual Meeting of Vision Sciences Society* (in its 13th year) serve as a measure of how far the field has come and what consistently stays unexplored. With many of the theories already presented by various researchers, it is now high time to integrate all the efforts from various communities of machine learning, neuroscience and physiology to reassure the objectives of learning efficient visual representations in machines. With joint efforts, one may discover the explanation for the underlying factors of image understanding in the visual cortex, as well as initiate significant debate on those areas where theories from different disciplines mismatch. It is anticipated that future advances in brain signal measurement and the development of more sophisticated encoding models of visual information will lead to a better apprehension of the complete neural model of human visual system, thus paving a way for human competitive object recognition systems.

REFERENCES

Aggarwal, J., Ghosh, J., Nair, D., and Taha, I. (1996). A comparative study of three paradigms for object recognition - bayesian statistics, neural networks and expert systems. In *Image Understanding: A Festschrift for Azriel Rosenfeld*, pages 241–262. Society Press.

Amit, D. and Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral corte. *Cereb. Cortex*, 7:237–252.

Bell, A. and Sejnowski, T. (1997). The 'Independent Components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338.

Biederman, I. (1986). Human image understanding: recent research and a theory. In *Second workshop on Human and Machine Vision II*, pages 13–57.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147.

Biederman, I. and Cooper, E. (1991). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, 23(3):393 – 419.

Bileschi, S. (2006). *Streetscenes: Towards Scene Understanding in Still Images*. PhD thesis, MIT.

Brunel, N. (2000). Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *Comput. Neurosci.*, 8:183–208.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, (ECCV)*, pages 1–22.

Csurka, G. and Perronnin, F. (2011). Fisher vectors: Beyond bag-of-visual-words image representations. In *Computer Vision, Imaging and Computer Graphics. Theory and Applications*, volume 229, pages 28–42.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893.

Deco, G. and Rolls, E. (2006). Decision-making and webers law: a neurophysiological model. *European Journal of Neuroscience*, 24:901–916.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *JMLR*, 11:625–660.

Fukushima, K. (1988). Neocognitron: a hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2):119–13.

Fulcher, E. (2003). *Cognitive Psychology*. NY, 1st edition.

Grimson, W. (1981). A computer implementation of a theory of human stereo vision. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 292(1058):217–253.

Hertz, J., Lerchner, A., and Ahmadi, M. (2004). Mean field methods for cortical network dynamics. pages 71–89. Springer-Verlag.

Hinton, G., Dayan, P., Frey, B., and Neal, R. M. (1995). The wake-sleep algorithm for self-organizing neural networks. *Science*.

Hinton, G. and Nair, V. (2009). 3D object recognition with Deep Belief Nets. In *NIPS*.

Hinton, G., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for Deep Belief Nets. *Neural Computation*, 18:1527–1554.

Hinton, G. and Salakhutdinov, R. (2009). Semantic hashing. *Approximate Reasoning*, 50(7):969–978.

Hubel, D. and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Physiology*, 160:106–154.

Hubel, D. and Wiesel, T. (1965). Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. *Neurophysiology*, 18:229–289.

Jaakkola, T. and Haussler, D. (1998). Exploiting generative models in discriminative classifiers. In *NIPS*, pages 487–493.

Karklin, Y. and Lewicki, M. (2005). A hierarchical Bayesian model for learning non-linear statistical regularities in non-stationary natural signals. *Neural Computation*, 17(2):397–423.

Karklin, Y. and Lewicki, M. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457(2):83–86.

Kreiman, G. (2008). Biological object recognition. *Scholarpedia*, 3(6):26–67.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *NIPS*.

Lades, M., Vorbruggen, J., Buhmann, J., Lange, J., Malsburg, C., Wurtz, R., and Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *Computers*, 42(3).

Larochelle, H. and Bengio, Y. (2008). Classification using discriminative restricted boltzmann machines. In *ICML*.

Leibe, B. and Schiele, B. (2003). Interleaved object categorization and segmentation. In *BMVC*, pages 759–768.

Li, T., Mei, T., and Kweon., I. (2008). Learning optimal compact codebook for efficient object categorization. In *IEEE Workshop on ACV*, pages 1–6.

Linsker, R. (1992). Local synaptic learning rules suffice to maximise mutual information in a linear network. *Neural Computation*, 4:691–702.

Logothetis, N. and Sheinberg, D. (1996). Visual object recognition. *Annual Review Neuroscience*, 19:577–621.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *ICCV*, volume 2, pages 1150–1157.

Lowe, D. (2004). Distinctive image features from Scale-Invariant keypoints. *IJCV*, 60(2):91–110.

Marr, D. and Nishihara, H. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. 200(1140):269–294.

Marr, D. and Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 204(1156):301–328.

Matas, J. and Obdrzalek, S. (2004). Object recognition methods based on transformation covariant features. In *EUSIPCO*.

Murase, H. and Nayar, S. (1995). Visual learning and recognition of 3-d objects from appearance. *Computer Vision*, 14(1):5–24.

Nevatia, K. and Binford, T. (1973). Structured descriptions of complex objects. In *IJCAI*, pages 641–647.

Nishimoto, S., Vu, A., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646.

Ojala, T., Pietikainen, M., and Harwood, D. (1994). Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *ICPR*, volume 1, pages 582–585.

Parikh, D. and Zitnick, C. (2010). The role of features, algorithms and data in visual recognition. In *CVPR*, pages 2328 –2335.

Perrett, D. and Oram, M. (1993). Neurophysiology of shape processing. *IVC*, 11:317–333.

Perronnin, F. and Dance, C. (2007). Fisher kernels on visual vocabularies for image categorization. In *CVPR*, pages 1–8.

Perronnin, F., Snchez, J., and Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *ECCV*.

Pinto, N., Cox, D., and DiCarlo, J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology*, 41(1).

Ponce, J., Berg, T., Everingham, M., Forsyth, D., Hebert, M., Lazebnik, S., Marszalek, M., Schmid, C., Russell, B., Torralba, A., Williams, C., Zhang, J., and Zisserman, A. (2006). Dataset issues in object recognition. volume 4170, pages 29–48. Springer Verlag.

Ramanan, A. and Niranjan, M. (2010). A One-pass Resource-Allocating Codebook for patch-based visual object recognition. In *MLSP*.

Reichert, D., Series, P., and Storkey, A. (2011a). Hallucinations in charles bonnet syndrome induced by homeostasis: a deep boltzmann machine model. In *NIPS*, volume 23, pages 2020–2028.

Reichert, D., Series, P., and Storkey, A. (2011b). A hierarchical generative model of recurrent object-based attention in the visual cortex. In *Artificial neural networks (ANN)*, ICANN, pages 18–25.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025.

Rolls, E., Loh, M., Deco, G., and Winterer, G. (2008-2009). Computational models of schizophrenia and dopamine modulation in the prefrontal cortex. *Nature Rev. Neurosci.*, 9(9):696–709.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Sanchez, J. and Perronnin, F. (2011). High-dimensional signature compression for large-scale image classification. In *CVPR*, pages 1665 –1672.

Sanchez, J., Perronnin, F., Mensink, T., and Verbeek, J. (2013). Image Classification with the FV: Theory & Practice. Technical Report RR-8209, INRIA.

Sejnowsky, T. (1976). On global properties of neuronal interaction. *Biol. Cybern*, 22:85–95.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., and Poggio, T. (2007a). A quantitative theory of immediate visual recognition. *Progress in Brain Research*, 165:33–56.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007b). Robust object recognition with cortex-like mechanisms. *PAMI*, 29:411–426.

Sugase, Y., Yamane, S., Ueno, S., and Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, pages 869–873.

Tarr, M. and Blthoff, H. (1998). Image-based object recognition in man, monkey and machine. *Cognition*, 67(1-2):1–20.

Taylor, G. and Hinton, G. (2009). Factored conditional Restricted Boltzmann Machine for modeling motion style. In *ICML*.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J., Lebihan, D., and Dehaene, S. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimg.*, 33(4):1104–1116.

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381:520–522.

Thulasiraman, K. and Swamy, M. (1992). *Graphs: Theory and Algorithms*. John Wiley & Sons, Inc., NY.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Cognitive Neuroscience*, 3(1):71–86.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume 1, pages I–511 – I–518.

Wallis, G. and Rolls, E. (1996). A model of invariant object recognition in the visual system. *Prog. Neurobiol.*, 51:167–194.

Wertheimer, M. (1938). *Laws of organization in perceptual forms*. W. Ellis, W (Ed. & Trans.), London: Routledge & Kegan Paul(Original work published in 1923).

Wilson, H. and Cowan, J. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, 12(1):1–24.

Wu, J. and Rehg, J. (2011). Centrist: A visual descriptor for scene categorization. *PAMI*, 33(8):1489–1501.

Zhang, J., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: a comprehensive study. *Computer Vision*, 73:213–238.