Memory–Prediction Framework for Pattern Recognition: Performance and Suitability of the Bayesian Model of Visual Cortex

Saulius J. Garalevicius

Department of Computer and Information Sciences, Temple University
Room 303, Wachman Hall, 1805 N. Broad St., Philadelphia, PA, 19122, USA
sauliusg@gmail.com

Abstract

This paper explores an inferential system for recognizing visual patterns. The system is inspired by a recent memory-prediction theory and models the high-level architecture of the human neocortex. The paper describes the hierarchical architecture and recognition performance of this Bayesian model. A number of possibilities are analyzed for bringing the model closer to the theory, making it uniform, scalable, less biased and able to learn a larger variety of images and their transformations. The effect of these modifications on recognition accuracy is explored. We identify and discuss a number of both conceptual and practical challenges to the Bayesian approach as well as missing details in the theory that are needed to design a scalable and universal model.

Introduction

For decades most artificial intelligence researchers tried to build intelligent machines that did not closely model the actual architecture and processes of the human brain. One of the reasons was that neuroscience provided many details about the brain, but an overall theory of brain function that could be used for designing such models was conspicuously lacking.

A recently published memory-prediction theory (Hawkins and Blakeslee 2004) offers a large-scale framework of the processes in the human brain and invites computer scientists to use it in their quest of machine intelligence. The first published attempt to model the theoretical framework (George and Hawkins 2005) can be considered as the starting point of the work described here. This paper describes experiments with the proposed Bayesian approach as well as explores the benefits, limitations and various extensions of the model.

The memory-prediction theory focuses on the functioning of the human neocortex. A hierarchical network structure guides the functioning of each region in the cortex. All regions in the hierarchy perform the same basic operation. The inputs to the regions at the lowest levels of the cortical hierarchy come from our senses and are represented by spatial and temporal patterns.

The neocortex learns sequences of patterns by storing them in an invariant form in a hierarchical neural network. It recalls the patterns auto-associatively when given only partial or distorted inputs. The structure of stored invariant representations captures the important relationships in the world, independent of the details. The primary function of the neocortex is to make predictions by comparing the knowledge of the invariant structure with the most recent observed details.

The regions in the hierarchy are connected by multiple feedforward and feedback connections. Prediction requires a comparison between what is happening (feedforward) and what you expect to happen (feedback). Each region is a collection of many smaller subregions that are only connected to their neighbors indirectly, through regions higher up the hierarchy. Each region learns sequences, develops "names" (invariant representations) for the sequences it knows and passes these names to the next region higher in the hierarchy. As a cortical region learns sequences, the input to the next region changes and the higher region can now learn sequences of these higher-order objects.

The remainder of this paper is organized as follows. The architecture of the original Bayesian model is described and the results of experiments with it are presented. Then the major drawbacks of this approach are identified and possible improvements are explored. The paper concludes with a discussion about the suitability and prospects of the memory-prediction framework and its Bayesian model for developing better performing pattern recognition systems.

Architecture of the Bayesian Model

A test application for recognizing visual patterns in black-and-white images was developed in C++ environment that allows to experiment with the Bayesian model. The original implementation was guided by the description of this model (George and Hawkins 2005) as well as prototype materials and sample images posted online by its authors.

The Bayesian model mirrors the large-scale biological structure of cortical regions and subregions connected in a functional hierarchy, but implements the functions of a subregion using Bayesian inference algorithms, not as a collection of columns and neurons. However, a fine map-

ping of these algorithms to the cortical anatomy has been found (George and Hawkins 2005).

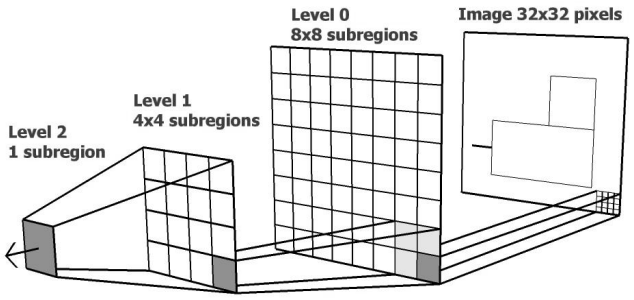

Figure 1. *Tree-shaped hierarchy. Each subregion in level 0 receives image fragment of size 4x4 pixels. Each subregion in level 1 receives input from 4 children in level 0. A single subregion in level 2 receives input from all level 1 subregions.*

The hierarchical tree-shaped structure (Figure 1) contains several subregions at each level, with the bottom-level subregions tiling the visual field and the single top subregion outputting the result of image recognition. Each subregion in level 0 remembers in a set B all input patterns with their percentage of occurrences in the training data greater than a threshold t. Then every pattern in B can be uniquely represented by its index k, which is called the "name" of the pattern and is produced as an output to level 1. Thus a subregion in level 1 is able to learn the most frequent simultaneously occurring combinations of patterns observed in its child subregions. Finally, the "name" (index k) of the active pattern in the parent subregion is fed back to its child subregions. This allows the subregions in level 0 to form a conditional probability distribution matrix that encodes a probability of observing known patterns given the patterns observed by the parent subregion. This learning process can be repeated for an arbitrary number of hierarchy levels and results in a tree-shaped Bayesian network. Therefore, recognition of any given image can be viewed as finding the set of known patterns at each subregion that best explains the image in a probabilistic sense. This is achieved using Pearl's Bayesian belief revision algorithm (Pearl 1988).

The memorized patterns of level 0 in the original model are predefined and consist of 139 primitives of commonly observed inputs (lines, corners, etc. in every possible position) together with the class index (meaning "vertical line", "top-right corner", etc.) for each primitive. The visual input to a subregion in level 0 is compared to the predefined patterns and the pattern with the smallest Hamming distance from the input is selected. Then the class index of this pattern is outputted as the "name" of the observed input. Subregions in level 1 memorize all observed patterns coming from child subregions (t = 0). The top level region has a fixed number of learned image categories as the output.

Performance of the Original Model

The model was trained with 91 categories of images (black-and-white line art symbols), containing 2 examples each (the "train" set). A person unrelated to this work generated the same number of testing examples by drawing reasonable, human-recognizable copies of the training images (the "test" set). The recognition accuracy recorded is the percentage of correctly recognized images after training. The model was tested using a variable number of saccades (eye movements) assuming that the accuracy should increase with more saccades. Eye movements were implemented by shifting the whole input symbol diagonally by one pixel per saccade and "bouncing" it at the sides of the visual input field.

Experiments showed that recognition accuracy of moving images is significantly better than that of flashed images, which agrees with the corresponding feature of human visual perception. The accuracy increases with the number of saccades used, however, the benefit of each additional saccade generally becomes less and less pronounced.

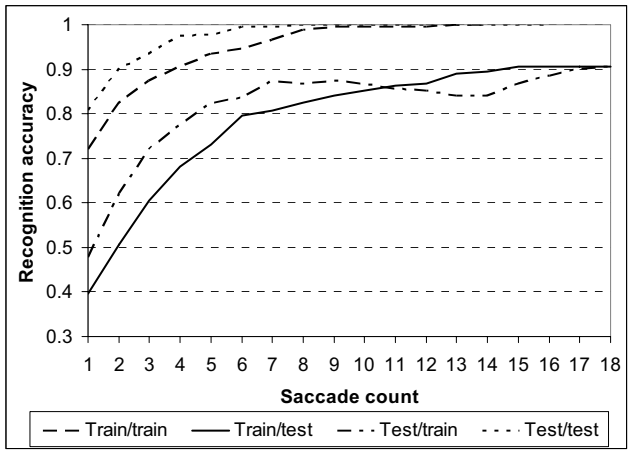

Figure 2. *Recognition accuracy of the original model*

Figure 2 shows the results of experiments with the "train" and "test" sets. The first set mentioned in the legend is used for training and the second one for testing. The model learns to recognize the original training images with 100% accuracy (after a certain number of saccades) when the same set is used for both training and testing. However, the recognition accuracy for unseen images is much lower. The "train/test" curve achieves a maximum accuracy of 90.6% after 15 saccades. This is significantly less than the 97% accuracy claimed by (George and Hawkins 2005); the authors likely used a smaller subset of training categories to achieve that result. This curve will be used here as a benchmark for all future performance comparisons.

The "train" set consisted of idealized images, where all examples were combinations of straight vertical and hori-

zontal lines, their intersections and right angles. As stated above, all these primitives are also stored in the memory of level 0, so the original set is specifically tailored to match the knowledge in the lowest region. Therefore it was useful to try training the model with less regular, haphazardly hand drawn "test" set. More variety in the training data requires the model to memorize more combinations of patterns which slows it down considerably. However, the same variety allows it to recognize both seen and unseen images with greater average accuracy, as evidenced by "test/test" and "test/train" curves in Figure 2. The latter curve also exhibits more complex non-monotonic behavior: the accuracy decreases at saccades 9 to 14, then increases again. Although this behavior is not desirable, it reflects the limitations of this model when trained with real world data as opposed to a designed set of regular examples.

Modifications of the Model

We modified the original model in a number of ways. The goal was to explore possible generalizations, bring the model closer to the memory-prediction theory, and see if a more faithful implementation of theoretical concepts leads to an improvement in recognition accuracy.

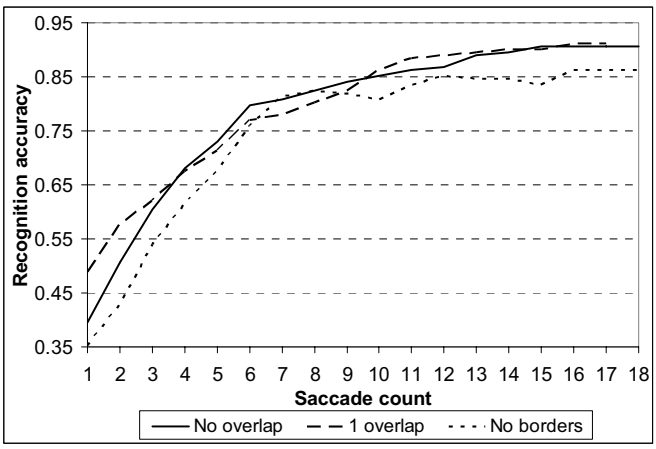

Figure 3. *Recognition accuracy with overlapping subregions and shifting beyond borders*

Overlapping Subregions

The memory-prediction theory states that adjacent subregions overlap within the same region of the brain. It is assumed that overlap leads to greater continuity and interrelationship between neighboring subregions thus improving predictive capabilities.

To test this assumption, the subregions in the bottom level were changed to receive partially overlapping inputs. This increases the size of regions: for 1 overlapping pixel we have 10x10 subregions in level 0 and 5x5 subregions in level 1. Despite the above assumption, testing showed that the overlap does not result in a definite improvement (Figure 3). Although the performance is better at 1-3 saccades, it fluctuates with increasing number of saccades and levels off at about the same accuracy as with no overlap. Therefore it cannot be concluded that overlap brings any meaningful improvement to the model, although it uses more memory and slows the processing considerably. It is evident that overlapping between subregions of the brain is much more complex. We could probably make better use of overlapping by including it in all hierarchy levels and by using more gradual transitions between neighboring subregions.

Recognizing Partial Images at Boundaries

Overall, the model exhibits a robust auto-associative capability and is usually able to recognize an image with some missing parts. One shortcoming of the original model is that each training image is moved only within the 32x32 pixel area without crossing its borders. This helps improve recognition of whole images, but it leads to inconsistent results when recognizing partial images. For example, the model recognized one half of the original "dog" image in just 1 saccade when the "half dog" was presented in the middle of the visual field. However, when the same image was shifted to the side of the visual field, the model failed to recognize it, because this pattern has never been presented in this position during the training. In fact, biological vision can recognize partial objects independent of their placement in the visual field.

To address this problem, an alternative training sequence was implemented. It moves the training image to all positions including those where only a part of the image appears in the visual field. As expected, this resulted in longer training time and decreased overall recognition accuracy by about 5% using varying number of saccades (Figure 3). However, now the model was able to recognize the "half dog" image shifted to the side as easily as that in the middle of the visual field. This type of learning better illustrates actual predictive capabilities of the model by using a more realistic training method.

Unified View of a Learning Region

As mentioned, the bottom region of the hierarchy does not function as a regular region in the original model. Instead, it "knows" a table of every possible vertical or horizontal line, corner, line intersection and similar primitives. It is assumed that these primitives are the most frequently observed patterns, which may or may not be the case depending on the input data. A subregion in level 0 finds the primitive closest to the input by selecting the smallest Hamming distance between them. Then it outputs the index of the category of the found primitive (not the index of the primitive itself). The bottom region uses only 14 different categories for output. So the bottom subregions communicate information such as "I see a vertical line" or "I see a top-right corner" to the higher level, thus performing

predefined classification of inputs. The problem is that this is a designed process, i.e. it is not proven that a region (or several regions) could effectively learn to communicate such information from input data.

The images in the "train" set intentionally consist of combinations of the primitives stored in level 0, since this matching helps learn the examples and recognize similar ones reasonably well. When the model is trained using our "test" set, this set also consists of similar patterns and therefore the model achieves comparable recognition performance. However, the original model cannot learn to effectively recognize any type of images (for example, images containing filled regions, diagonals, wider lines, etc. are recognized very poorly). The performance of the model is biased towards recognizing images consisting of similar components as those stored in the bottom region.

According to the memory-prediction theory, all regions of the brain perform the same operation and can effectively learn based on any type of input. Hence designing the model according to these principles will both make it more universal and unveil its real, less biased capabilities.

To achieve these goals the lowest region of the hierarchy was modified so that it no longer stores a predefined group of primitives. Instead, it memorizes the patterns based only on the training data. Thus all regions function according to identical principles and have no predetermined knowledge.

After memorizing all the patterns observed in the "train" set, the alphabet of the bottom region contains 174 values that it communicates to its parent region, as opposed to just 14 categories before the change, a 12 times increase. The subregions in the parent region in turn observe more possible combinations of these values, increasing the actual used memory more than 7 times. Hence the conditional probability distribution matrices occupy 91 times more memory than before in each subregion of level 0. It is clear that in order to prevent a combinatorial explosion as we go towards the higher levels in more realistic systems we need to tackle the issue of forgetting less important memories.

Forgetting: Memory Usage vs. Accuracy

The human brain does not memorize all patterns, just the ones that are most frequently observed in a dynamic environment. In our controlled testing environment we can model this process by calculating the frequency of patterns in the training data and discarding less frequent patterns.

The learning algorithm was modified to determine the frequency of observed patterns and forget rarely observed ones in each hierarchy level. To achieve this, the learning process was made incremental and uses several passes to train each hierarchy level. After the level being trained memorizes its patterns, we discard the patterns with observed frequency below a specified threshold. Since that region will no longer be memorizing new patterns, contextual information about the most frequent patterns can now be fed back to its child region to form conditional probability distribution matrices. If the region encounters an unfamiliar pattern in the future, it selects the most similar pattern from its memory as the output. The process continues

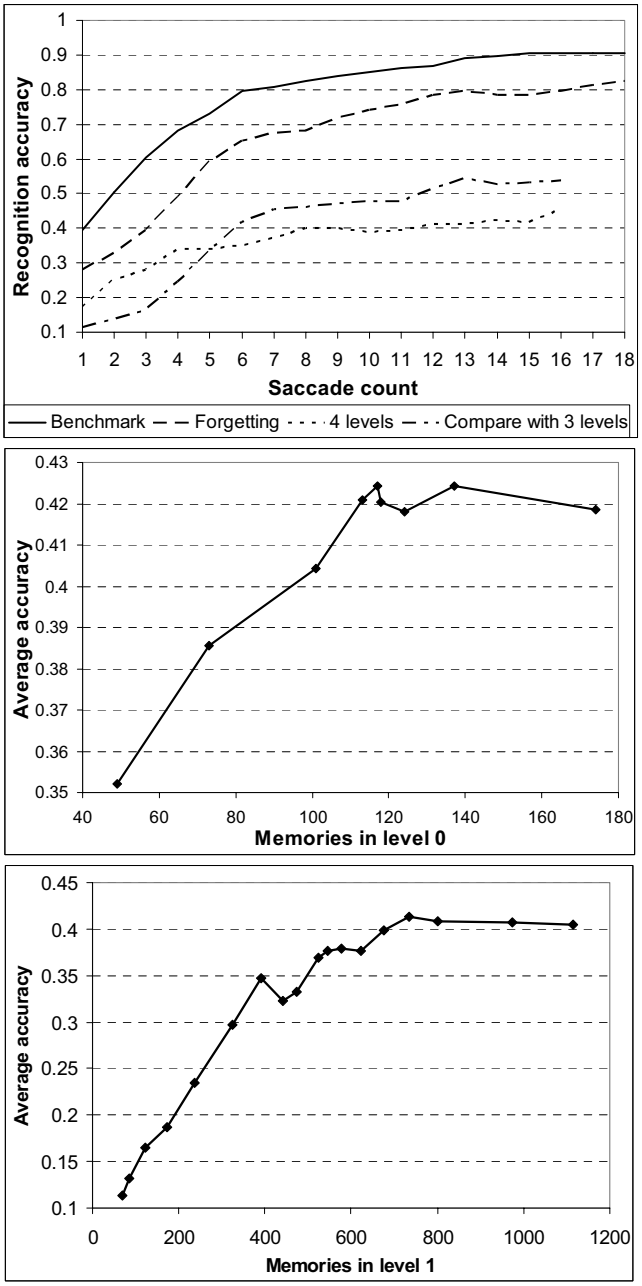

Figure 4. *Top: recognition performance using the unified view of a learning region and forgetting; performance using additional hierarchy level. Middle and bottom: dependency between the weighted average accuracy and the number of memories after forgetting in levels 0 and 1 respectively.*

training regions one by one, until the top region has learned its patterns. By adjusting the forgetting thresholds, we can manage memory issues while trying to achieve reasonably good recognition performance.

Experiments were performed using the new model to explore how various forgetting thresholds affect recogni-

tion accuracy. One of the best results that can be achieved with reasonable memory usage is shown in the top graph of Figure 4 as the "Forgetting" curve. This experiment had 101 memories in level 0 and 800 memories in level 1 after forgetting. As we see, the accuracy is lower than that of the original model by about 10%. Even if we use the same number of memories in level 0 after forgetting as there were predefined primitives in the original model (139 primitives), we do not get a comparable accuracy (we also use moderate forgetting in level 1 to manage memory issues). This strongly suggests that this model is not able to learn to classify its most frequently observed patterns as effectively as the predefined classification provided in the original model. (To actually prove this claim we would need to eliminate forgetting from level 1.) Hence the real capabilities of the model that does not use any additional predefined knowledge seem to be significantly lower.

The influence of forgetting threshold upon the achievable accuracy was also explored. It appears that we can forget a lot of insignificant memories without a significant drop in accuracy. Of course, there is a limit where accuracy starts to degrade sharply. The middle and bottom graphs in Figure 4 show the dependency between the number of memories after forgetting in levels 0 and 1 and the weighted average of recognition accuracy for saccade counts from 1 to 4 (accuracy using a larger number of saccades gets more weight). In each experiment the forgetting threshold was changed in only one level while keeping it constant in the other level. We see that there exists an optimal number of memories in a region where the recognition capability of the model is maximized. Even as the number of memorized patterns is further increased, the accuracy does not improve, but tends to become slightly lower compared to the optimal point. Therefore the forgetting strategy pays off especially when removing excessive memories that do not contribute to any additional gain in recognition accuracy.

Adding Additional Hierarchy Levels

The memory-prediction theory describes the cortical hierarchy with many regions. Each additional region forms higher-order knowledge and captures more complex invariances based on the inputs from its child region. All regions function according to the same principles and there is no theoretical limit to the number of hierarchy levels.

It is obvious that in order to perform complex tasks, any model has to be driven by the same principles and the hierarchy has to be scalable. Since in our current model all regions already perform the same basic operation, it is theoretically quite straightforward to add any number of additional regions without violating the overall architecture.

However, practical experiments with an additional hierarchy level showed that it is practically impossible to achieve any benefit from the expanded hierarchy due to the memory constraints. The top graph in Figure 4 shows the best result achieved using 4 hierarchy levels instead of 3 (the "4 levels" curve). Note that the forgetting threshold for level 1 is much lower than in the previous experiment: level 0 has 101 memories as before, while level 1 has only 174 memories. We can see from average accuracy analysis in Figure 4 that so little memories in level 1 result in very inaccurate output from that level to begin with. Since the added region depends solely on this output, it is no wonder that the overall result is so inaccurate. And even this relatively low number of memories in level 1 yields a whopping 26985 memories in our added level 2 before forgetting. As a comparison, the same number of memories in level 0 produces just 7798 memories in level 1 before forgetting. The effects of combinatorial explosion that were discussed above become increasingly pronounced higher in the hierarchy. Generally, forgetting allows limiting the number of memories to any manageable number; however, the lower levels in the expanded hierarchy had to forget so much that it degraded their performance. Therefore the addition of an extra hierarchy level did not improve the overall accuracy of the model.

An experiment was also performed using the same forgetting thresholds in the two bottom levels with and without the added additional level. It was designed to reveal if the extra level brings any benefit to the model, all other parameters being equal. The result is inconclusive, as can be seen from the top graph in Figure 4. The larger hierarchy performed better with smaller saccade count (up to 5 saccades), but the smaller hierarchy excelled with a larger saccade count.

Discussion and Conclusions

The model described here is an uncommon, biology-inspired approach to the pattern recognition problem. For a long time neural networks have been a popular approach for recognizing visual patterns. Although initially inspired by the microscopic structure of the human brain, neural networks usually have little similarity with the actual processes in the brain, other than both being constructed of many interconnected units (Hawkins and Blakeslee 2004).

This is just one of many possible models of the memory-prediction theory. This design repeats the biological hierarchy of regions and subregions, but simulates the detailed functioning of subregions using Bayesian belief revision techniques stemming from computer science. As such, this model is closer to an inferential system than to a dynamical system, although it shares many common ideas with traditional neural networks (Wang 2006). The inferential nature of the model eliminates or reduces some of the disadvantages of neural networks. In principle, the model could accept additional training categories even after the initial learning phrase is completed. It can provide several reasonable outputs for a single ambiguous input, while the learned internal knowledge can be interpreted with less difficulty by a human being. Finally, it offers greater promise of understanding what intelligence is by modeling the overall large-scale structure of human neocortex. Experiments with this probabilistic model showed that it recognizes scaled, translated, distorted and noisy images well,

the versatility not matched by neural networks and many other approaches (George and Hawkins 2005).

The reaction to the new memory-prediction theory has been mostly guarded. Many authors pointed out the incompleteness of the theory, suggested adding more essential features and a discussion has ensued (Feldman 2005, Perlis 2005, Taylor 2005, Hawkins 2005). A new technology called Hierarchical Temporal Memory is being developed based on the original probabilistic model by Numenta, Inc.; however, little details about it have been published to date.

Despite a number of advantages mentioned above, this research uncovered some theoretical and practical shortcomings of the probabilistic model. Although the learning process is not iterative (each image is seen only once), in practice it is still time consuming. The image has to be presented in all possible positions within the visual field during training to learn translation invariance. In comparison, the dorsal stream of the visual cortex uses observed visual information to guide much lesser number of saccades for examining an object. More accurate modeling of this intelligent examination of a given object is crucial for improving the performance of the model. The program also slows down considerably with increasing memory usage in regions.

This issue of memory usage is one of the main problems with this model. The number of possible combinations of outputs from child subregions rapidly increases as we go higher in the hierarchy. In addition, the model relies on large conditional probability distribution matrices for each subregion, compounding the memory growth. To address the former problem we are forced to use radical forgetting techniques, while the latter one may be somewhat mitigated by using storage techniques designed for sparse matrices. However, in order to maintain a practicable number of memories, the model has to forget so much knowledge that it severely deteriorates its performance, especially when adding additional hierarchy levels. Constructing a hierarchy with an arbitrary number of levels is essential for learning complex invariant representations, so a more effective learning and memory management approach is especially needed for making the model scalable and universal.

There are several key areas where the model fails to reflect the fundamental processes of the brain, as described by the memory-prediction theory. Despite (George and Hawkins 2005) emphasizing sequences of moving images, the regions do not store sequences in time, nor does the model learn to predict any movement. Instead, only single patterns are remembered during training, while image movement in the testing phase serves to poll different subregions in order to produce their combined prediction using Bayesian belief revision. Since storing and predicting time sequences is among the cornerstones of the theory, this capability should be included in future models.

It was also observed during experimentation that the model tends to classify an image by identifying many small details (line segments, corners, etc.) that were also present in the training example, but not by relying on the overall appearance or topology of the image. Hence some obviously vital elements of the test image are discounted and this leads to many of the observed classification errors. By contrast, human perception makes use of the large-scale topological structure, quickly detects most interesting parts of the image and guides the saccades to examine these parts more closely.

Overall, the memory-prediction theory provides a convincing large-scale framework of information processes in the neocortex that can be used as a foundation of many models. The probabilistic model described here proves to be a useful tool for gaining new insights about the details and performance of the systems based on this theory. However, the theory still lacks many crucial details needed to implement practical systems and the Bayesian model does not offer viable solutions to fill many of these gaps. Therefore the model suffers from both conceptual and viability problems. Nevertheless, it is to be expected that this new theoretical framework will mature and give rise to more accurate models, leading to significant achievements in our quest to build truly intelligent machines.

References

Feldman, J. A. 2005. On Intelligence as Memory. *Artificial Intelligence* 169(2005):181-183.

George, D., and Hawkins, J. 2005. A Hierarchical Bayesian Model of Invariant Pattern Recognition in the Visual Cortex. In Proceedings of the International Joint Conference on Neural Networks 2005. Montreal, Canada: International Neural Network Society.

Hawkins, J., and Blakeslee, S. 2004. *On Intelligence.* New York, NY: Times Books.

Hawkins, J. 2005. Response to Reviews by Feldman, Perlis, Taylor. *Artificial Intelligence* 169(2005):196-200.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems.* San Mateo, CA: Morgan Kaufmann Publishers.

Perlis, D. 2005. Hawkins on Intelligence: Fascination and Frustration. *Artificial Intelligence* 169(2005):184-191.

Taylor, J. G. 2005. Book Review. *Artificial Intelligence* 169(2005):192-195.

Wang, P. 2006. Artificial General Intelligence and Classical Neural Network. In Proceedings of the IEEE International Conference of Granular Computing, 130-135. Atlanta, Georgia: IEEE Computational Intelligence Society.