The Impact of Letter Classification Learning on Reading

Gale L. Martin
MCC
3500 Balcones Center Drive
Austin, Texas 78759
galem@mcc.com

Abstract

When people read, they classify a relatively long string of characters in parallel. Machine learning principles predict that classification learning with such high dimensional inputs and outputs will fail unless biases are imposed to reduce input and output variability and/or the number of candidate input/output mapping functions evaluated during learning. The present paper draws insight from observed reading behaviors to propose some potential sources of such biases, and demonstrates, through neural network simulations of letter-sequence classification learning that: (1) Increasing dimensionality does hinder letter classification learning and (2) the proposed sources of bias do reduce dimensionality problems. The result is a model that explains word superiority and word frequency effects, as well as consistencies in eye fixation positions during reading, solely in terms of letter classification learning.

Introduction

Models of word recognition and reading typically focus on processes that occur after letters are classified. They explain reading behaviors in terms of the processes that act on the outputs of letter detectors, rather than in terms of the processes that convert the image of a letter string to the corresponding outputs of the letter detectors. For example, Morton's Logogen model (Morton, 1969) focuses on word-level representations, and explains *word frequency effects* [1] in terms of lowered activation thresholds for word detectors. Similarly, McClelland & Rumelhart's (1981) Interactive Activation model focuses on associations and interactions between letter detectors and word detectors, and explains word frequency and *word superiority effects* [2] by proposing that these associations amplify the activation coming from letter detectors. The present paper departs from this tradition by proposing a model of letter classification learning. The model explains word frequency and word superiority effects, as well as certain regularities in eye fixation positions solely in terms of factors that determine letter classification learning.

The model was suggested by an interesting difference between human reading and machine-based Optical Character Recognition (OCR) systems (see Figure 1). Whereas OCR systems classify individual letters; human readers classify letter sequences. That is, people classify a sequence of as many as 8-13 characters within a single fixation (Rayner, 1979) and within such a fixation, the classification occurs in parallel (Reicher, 1969; Blanchard, McConkie, Zola & Wolverton, 1984).

One reason why this is an interesting difference is that if people classify a sequence of letters together, within essentially one operation, then we might expect that the familiarity of the letter sequence would impact letter classification operations, as well as subsequently occurring processes. In other words, we might expect general reading behaviors to be determined by letter classification processes, as well as processes that involve higher-level representations.

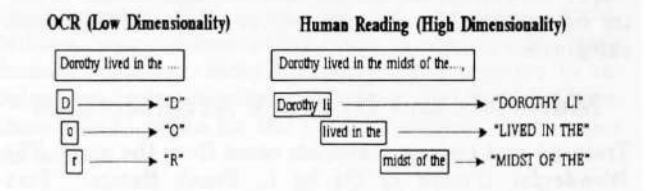

Figure 1: Character classification versus character sequence classification.

A second reason why the difference between OCR systems and human reading is interesting is that OCR systems are designed to classify single characters to minimize the so-called *curse of dimensionality* (Denker, et al, 1987). This machine learning principle predicts that high dimensionality (for example, large, high detail inputs) will cause classification learning to fail. If no constraints are placed on inputs, outputs, and input-output mapping functions, increased dimensionality leads to exponential increases in the number of different inputs, outputs, and mapping functions. Classification learning corresponds to approximating a particular mapping function by sampling from its population of input-output pairs. Learning will fail, with high dimensional inputs and outputs, because: (1) the system can not sample enough pairs to capture the full variability of inputs and outputs; thereby causing low generalization, and/or (2) it has insufficient capacity to describe even the sampled training pairs. Geman, Bienenstock & Doursat (1992) point out that the answer to this dilemma is not to put

[1] People identify high frequency words faster than low frequency words

[2] People identify letters within words and pronounceable non-words faster than they identify isolated letters and letters within unpronounceable non-words

one's energies into developing yet another new learning algorithm, because all classification learning algorithms will be subject to the curse. Instead they recommend developing an understanding of how to appropriately bias learning in given domains. Biasing corresponds to implementing *a priori* assumptions that rule out, or render less likely, some portion of the set of all possible inputs, outputs, and/or mapping functions.

This machine learning perspective, combined with the psychological evidence that people do learn to classify high dimensional images of letter sequences, suggests that it might be useful to view reading behaviors in terms of biases that make accurate letter classification possible. The model proposed here assumes that people would fail to learn to classify high dimensional letter sequences unless such learning was biased, and that specifying these biases may help explain a variety of reading behaviors. Three sources of bias, or methods for reducing learning complexity, are proposed here: (1) Limiting mapping functions to those based on position-invariant local feature detectors, and limiting the range of inputs and outputs by limiting the range of, or variability in, (2) eye fixation positions, and (3) allowed character sequences. This paper describes three experiments that support the model. Each consists of a set of backpropagation (Rumelhart, Hinton, & Williams, 1986) neural net simulations of letter classification learning, in which the inputs are images of individual letters or letter strings and the outputs are the corresponding letter categories.

Materials and Network Architectures

Training and testing materials came from the story *The Wonderful Wizard of Oz* by L. Frank Baum. Text line images were created from 120 pages of text (about 160,000 characters, 33,000 total words, or 2,600 different words), divided into 6 sets of 20 pages each. Each set was printed in 1 of 3 fonts, and in either all upper case characters or the original mix of lower and upper case (see Figure 2). Two of the three font types had variable-width characters, and one had constant-width characters. It was important to include variations in character widths because classifying letter sequences involves locating the relative positions of each character identified. Each image was labeled with the categories and horizontal positions of the letters depicted. Text line images were normalized with respect to height, but not width. Training and test sets contained an equal mix of the six font/case conditions. Two generalization sets were used, for test and cross-validation, and each consisted of about 14,000 characters. Training performance was measured by two metrics: (1) Asymptotic accuracy on the training data, and (2) amount of training required to reach asymptote. Generalization performance was measured by accuracy on test set, and on the cross-validation set.

The neural network architectures used here are extensions of the local receptive field, shared weight architectures (see Figure 3) used in some OCR systems (LeCun, et al, 1990; Martin & Pittman, 1991). In this earlier version, the input is the image of a single letter, and the output a vector representing the letter category. Hidden nodes receive input from a local region (e.g., a 6x6 area) in the layer below. Hidden layers are visualized as cubes, made up of separate planes. Hidden nodes within a plane share weights. Corresponding weights in the nodes' receptive fields are randomly initialized to the same value and updated by the same error, so that different hidden nodes within a plane learn to detect the same feature at different locations. Different feature detectors emerge from hidden nodes within different planes due to different random initializations. Output nodes are connected to all nodes in the previous layer, but not each other.

Dorothy lived in the midst of the great Kansas Prairies.
DOROTHY LIVED IN THE MIDST OF THE GREAT KANSAS PRAIRIES.
Dorothy lived in the midst of the great Kansas Prairies.
DOROTHY LIVED IN THE MIDST OF THE GREAT KANSAS PRAIRIES.
Dorothy lived in the midst of the great Kansas Prairies.
DOROTHY LIVED IN THE MIDST OF THE GREAT KANSAS PRAIRIES.

Figure 2: Samples of type font and case conditions

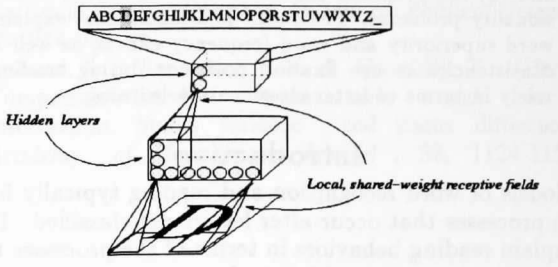

Figure 3: Local, shared weight architecture common in OCR systems.

The extension to an architecture that classifies character sequences is illustrated in Figure 4, for the case in which k, the number of to-be-classified characters, is equal to 4. The input window is expanded horizontally to cover k of the widest characters ("WWWW"). The image of a string of narrower characters will depict additional characters to the right, which the net must learn to ignore. Hidden layers are also expanded horizontally. Network capacity is described by the depth (the number of different feature detectors, or planes) and width of each hidden layer. Each output node represents a character category in one of the kth ordinal positions in the string. Networks were trained until the training set accuracy failed to improve by at least .1% across 5 passes through the training set. Nets were monitored for overfitting using the test set, but such overfitting never occurred.

This architecture biases learning by selectively reducing network capacity relative to that of a comparable globally connected net. The bias seems to be favorable, in that these nets could be trained with at least moderate success; whereas attempts to train globally connected nets on character sequence images failed miserably.

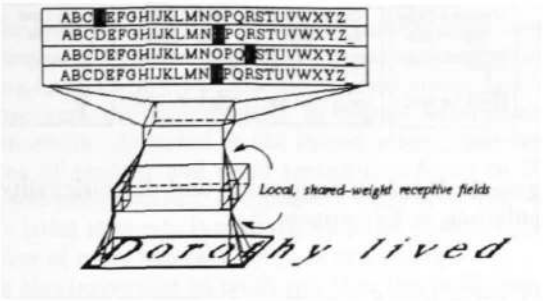

Figure 4: Net architecture for parallel character sequence classification, k=4 characters.

A limited claim can also be made that the biases imposed are similar to those imposed by mammalian visual systems. Like the local, shared-weight architecture, mammalian visual systems appear to use spatially local feature detectors that are replicated across the visual array (Hubel & Wiesel, 1979). There is also some very rough similarity between the oriented edge and bar detectors that emerge in both systems (Hubel & Wiesel, 1979; Martin & Pittman, 1991), as illustrated in Figure 5, which depicts some receptive fields that developed in first hidden-layer nodes with the local, shared-weight networks trained on letter images (Martin & Pittman, 1991). These receptive fields indicate that the corresponding feature detecting nodes discriminate on the basis of oriented edges and bars, but beyond this, any similarity to human or mammalian vision systems is unknown.

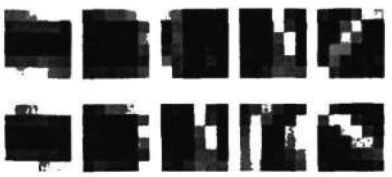

Figure 5: Feature detectors that emerged in OCR neural nets

Curse of Dimensionality Effects

The purpose of Experiment 1 was to test whether high dimensionality is associated with decreased training and generalization accuracy, even when learning is biased through the use of the local, shared-weight architecture. That is, the goal was to determine if classification learning becomes increasingly difficult as we move from the situation in which only single characters are classified, to that in which a letter sequence is classified. Four levels of dimensionality were examined (see Figure 6), ranging from a 20x20 input window, k=1; to an 80x20 input window, k=4. Input images were generated by starting the window at the left edge of the text line, with the first character centered 10 pixels from the left of the window, and then successively scanning to the right, pausing at each character position. Five different training set sizes

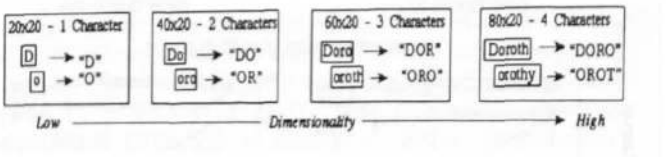

Figure 6: Four levels of input/output dimensionality used in the experiment.

were used (roughly 700 samples to 50,000), as well as a lower and higher capacity version of each network (15 vs 18 different feature detectors in each hidden layer). 40 different networks were trained, one for each combination of dimensionality, training set size and relative network capacity (4x5x2). Some nets required several months to train.

Accuracy is reported as percent of fields in which all characters were correctly classified. The results (see Figure 7) confirm the curse of dimensionality prediction that increasing dimensionality hinders both training and generalization. Increasing dimensionality lowers asymptotic accuracy achieved on the training set ($F(3,27) = 15.15, p < .01$), and increases the number of training passes required to reach the asymptote ($F(3,27) = 14.44, p < .01$). It also decreases generalization accuracy rates on both the test set ($F(3,27) = 33.9, p < .01$) and the validation set ($F(3,27) = 61.38, p < .001$). These results suggest that, even with the constrained architecture, high dimensionality leads to inadequate classification learning. Since human reading appears to involve even higher dimensionality than that modeled here, these results argue for the need for factors that reduce variance.

Constraints on Fixation Positions

One method for reducing the variance in to-be-classified images is to constrain the variability in fixation positions within a word. The input images in Experiment 1 were generated from fixations at each character position within a word, and thus were highly variable. However, people fixate most often at a *preferred viewing location*– slightly to the left of the middle of a word (Rayner, 1979). The non-randomness of eye fixation positions would have the effect of reducing image variability, and hence should aid classification learning. Accordingly, people do identify a word more quickly when the eyes are fixated near this location (O'Regan & Jacobs, 1992). Besides the benefit of consistency, the location may be optimal in that the average variability in the distance of characters from fixation is minimized when a point toward the middle of a word is fixated.

Experiment 2 used four different conditions to examine whether such consistent and optimal positioning reduces dimensionality problems. The *consistent and optimal positioning* condition used an 80x20 input window positioned with respect to the 3rd character of each word of 3 or more characters (see Figure 8)[3] and the net was trained to classify the first 4 characters in the word. The

[3]This is a simplification of the fixation position consisten-

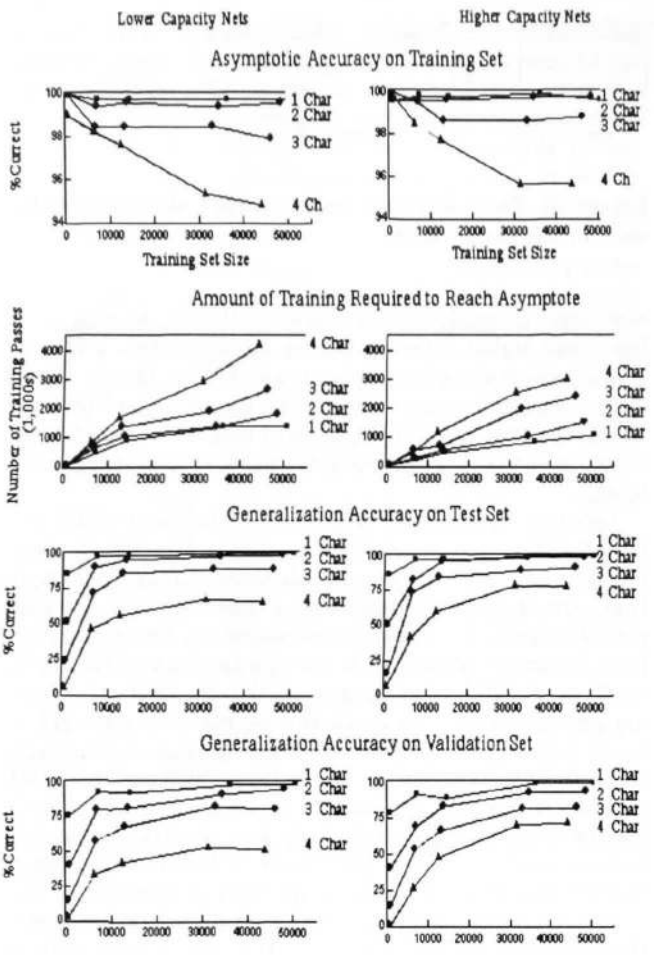

Figure 7: Impact of dimensionality on training and generalization.

consistent positioning only condition differed from this only in that positioning was with respect to the first character of a word. The *high dimensionality control* condition used the same input/output dimensionality, but the window was positioned at all character positions during training, and at the first character in a word during testing. The *low dimensionality control* used a 20x20 input window, with k=1, and the net trained and tested only on the first 4 characters. Four levels of training set size were used, with three replications of each training set size x window condition, resulting in 4 x 4 x 3 = 48 networks trained and tested. All networks employed 18 different feature detectors for each hidden layers.

The results support the value of both consistent and optimal positioning in reducing dimensionality problems (see Figure 9). The effects of positioning were significant with respect to asymptotic accuracy ($F(3,32) = 71.83, p < .001$), and generalization in both the test and validation sets ($F(3,32) = 861.9, p < .001$; and

cies in human reading, which are better described in terms of a probability distribution, the mean of which falls toward the center of a word.

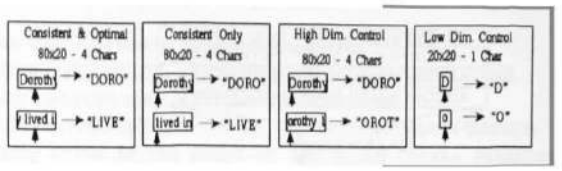

Figure 8: Window positioning and dimensionality manipulations in Experiment 2

$F(3,32) = 1022.6, p < .001$). Subsequent t-tests revealed that the high dimensional control condition networks did worse than the nets in the other three conditions across all three of the metrics which resulted in significant analysis of variance results. Moreover, the consistent and optimal positioning networks yielded better asymptotic training and generalization accuracies than those in the consistent positioning only condition, and better than or equivalent to those in the low dimensionality control condition. The consistent positioning only condition generally performed better than the high dimensionality control. These results support the value of both consistent positioning and optimal positioning in reducing the negative effects of dimensionality on classification learning.

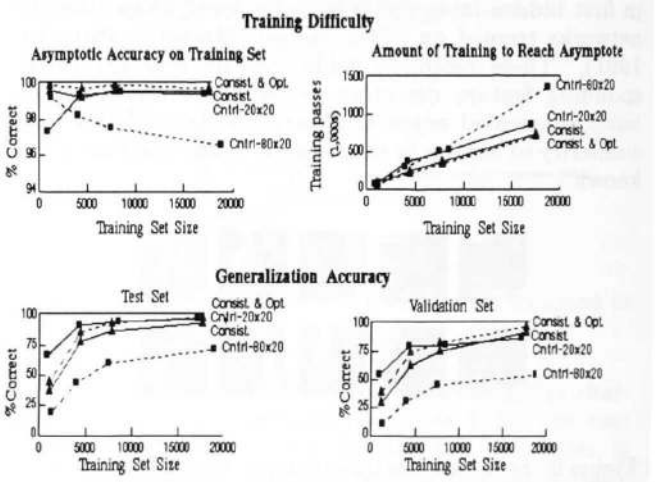

Figure 9: Impact of consistent & optimal window positions.

Character Sequence Regularities

Another factor that may reduce the complexity of classification learning stems from the fact that English words constitute a subset of all possible letter sequences, and of this subset, not all sequences have an equal likelihood of being encountered. People appear to take advantage of such constraints when they read, in that they are better at reading familiar, as compared to unfamiliar, letter sequences. They identify letters within words faster than letters within non-words, and letters within pronounceable non words faster than letters within random character strings (Baron & Thuston, 1973; Reicher, 1969), and they identify high frequency words more quickly than low frequency words (Solomon & Postman, 1952). Note

that the reading system could take advantage of the non-randomness of letter sequences at the letter classification learning level, to improve classification accuracy; and/or at subsequent processing levels, to correct letter classification errors. As noted in the introduction, most conceptions of reading and word recognition focus on the latter processing levels. The present focus is on the impact of letter sequence familiarity on letter classification, exclusive of more abstract levels of processing.

It is also important to point out that the feedforward networks used here can not reflect processing times directly, since each forward pass in the net takes the same amount of time. Therefore, in modeling word superiority and word frequency effects, it is assumed that lower accuracy rates translate to slower performance, either because lower accuracy rates would require additional, time-consuming post-processing mechanisms to correct classification errors or because, in an interactive activation network, the reduced activation associated with less certain responses would be reflected in longer times to reach thresholds of activation.

Experiment 3 involved seeing if the three best consistent and optimal positioning nets from Exp. 2 exhibit human-like word superiority and frequency effects in the sense of exhibiting lower accuracy for the less familiar letter sequences. The control condition used the nets trained in the low dimensional control condition to distinguish between effects due to individual letter familiarity and effects due to letter sequence familiarity. New text images were created to produce the following sets or conditions. The *word* set had 30 4-letter words, drawn from the Oz text, of which 15 occurred very frequently in the text (e.g., SAID), and 15 occurred infrequently (e.g., PAID). The *pronounceable non-word* set had 30 4-letter pronounceable non-words (e.g., TOID). The *random non-words* had 30 4-letter random strings (e.g., SDIA). The *alternating case words* used the word set but the letters were printed in AlTeRnAtInG cases. This latter set was created to see if the nets exhibit human-like behavior in being able to read despite such manipulations (McClelland, 1976). Word superiority results were analyzed in terms of a split-plot analysis of variance with letter sequence type and dimensionality as factors, and associated t-tests. Word frequency results were analyzed with t-tests.

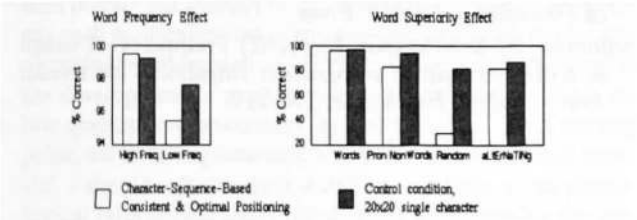

Figure 10: Sensitivity to word frequency and character sequence regularities

As shown in Figure 10, the consistent and optimal positioning nets mimic word frequency and word superiority effects. In the case of word superiority effects, significant main effects were found for letter sequence type ($F(3,12) = 181.8, p < .001$) and dimensionality ($F(1,4) = 77.4, p < .001$). The interaction was also significant ($F(3,12) = 80.2, p < .001$). Paired comparison tests confirmed the advantage for words over the other letter sequence types; and for pronounceable non-words and aLtErNaTiNg case words over random non-words. Letter classification accuracy remains high in spite of the words being printed in alternating cases, which is presumably due to the local, shared-weight architecture biasing the system toward local, rather than word-level, feature detectors. The consistent and optimal positioning nets also showed a tendency to classify high frequency words more accurately than low frequency words ($p = .05$). These results are interesting because they support the notion that word superiority and word frequency effects can be explained without reference to higher levels of processing.

Discussion

More generally, the results presented here support the value of viewing reading behaviors in terms of biases that make it possible to learn to accurately classify letters. Experiment 1 demonstrated that classification accuracy drops dramatically with increases in the size of the to-be-classified image and the number of to-be-classified letters. Experiment 2 demonstrated that these negative effects of dimensionality can be offset, at least to some extent, through the use of a simplified form of the consistent fixation positions used in human reading. Experiment 3 demonstrated that the letter classification learning system exhibited word superiority and word frequency effects similar to those of human readers, even though there were no higher level representations such as words or phonological codes in the system.

The results also raise several issues for discussion. One of these is the question of whether or not additional processing levels also determine word frequency and word superiority effects. It might be argued on the grounds of parsimony that there is no need to model additional processing mechanisms, since the relatively low level, classification mechanism can account for the findings. However, it seems more reasonable to assume that learning acts at multiple levels because letter classification processes are likely to need all the help they can get. Current extensions of the work reported here involve expanding the input window to cover 8 or more letters, as well as requiring the network to learn to classify 8 or more letters. This work indicates that classification accuracy drops considerably with such extensions, and therefore it seems reasonable to propose that word-level or phonological-level coding would still play a critical role in improving letter classification accuracy.

The results also raise the question of whether or not factors that determine letter classification learning also determine reading disabilities and developmental stages of reading. The present work demonstrates the importance of consistencies in eye fixation positions. Some reading problems are associated with reduced

input/output dimensionality, as measured by perceptual span (Rayner, 1986; Rayner, et al., 1989) and with irregular eye fixation patterns (Rayner & Pollatsek, 1989). Such irregularities would increase input/output variability, and hence reduce the dimensionality at which high accuracy levels could be maintained. This pattern supports the relevance of letter classification learning factors to reading disabilities and developmental differences. Perceptual and classification processes have sometimes been discounted as causes of reading disabilities on the grounds that reading disabilities and developmental differences become more apparent with more difficult content. Content factors have traditionally been associated with processes beyond letter classification. The present results suggest that this assumption warrants further consideration since a factor that is often associated with content difficulty–word frequency–was shown to impact classification accuracy.

A third issue pertains to the question of why the human reading system doesn't avoid all of these dimensionality problems by taking the same approach chosen by developers of optical character recognition systems–classifying individual letters rather than letter sequences. One possibility is that the brain can't easily separate small (letter sized) individual parts of an image, classify each and retain the original order of the images to infer letter sequence information, and so it is forced into dealing with the high dimensionality. In this case, the human mechanisms would be less optimal than the corresponding machine-based mechanisms. Alternatively, it is possible that incorporating letter sequence familiarity at multiple stages of processing, as is possible in the current system, would lead to higher overall accuracy rates. In this case, the human mechanisms would be superior to the machine-based mechanisms.

The final issue pertains to the direction of future research. As noted earlier, the current focus is on developing a model that can at least partially classify between 8 to 13 letters within a single "fixation." Once this model has reached some degree of stable performance, the focus will shift to incorporating additional aspects of reading. One of these aspects is the control of eye movements. Previous work (Martin, Rashid & Pittman, 1993) indicates that it is possible to train networks to generate ballistic and corrective saccades to navigate along a path of text. Other aspects include integrating the information obtained from successive fixations, and using word-level information to improve classification accuracy.

References

Blanchard, H., McConkie, G., Zola, D., & Wolverton, G. (1984) Time course of visual information utilization during fixations in reading. *Jour. of Exp. Psych.: Human Perc. & Perf., 10*, 75-89.

Denker, J., Schwartz, D., Wittner, B., Solla, S., Howard, R., Jackel, L., & Hopfield, J. (1987) Large automatic learning, rule extraction and generalization, *Complex Systems, 1*, 877-933.

Geman, S., Bienenstock, E., and Doursat, R. (1992) Neural networks and the bias/variance dilemma. *Neural Computation, 4*, 1-58.

Hubel, D. & Wiesel, T. (1979) Brain mechanisms of vision. *Sci. Amer., 241*, 150-162.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1990) Handwritten digit recognition with a backpropagation network. In *Adv. in Neural Inf. Proc. Sys. 2*, D. Touretzky (Ed) Morgan Kaufmann.

Martin, G. L. & Pittman, J. A. (1991) Recognizing hand-printed letters and digits using backpropagation learning. *Neural Computation, 3*, 258-267.

Martin, G. L., Rashid, M., & Pittman, J. A. (1993) Integrated segmentation and recognition through exhaustive scans or learned saccadic jumps. In *Advances in Pattern Recognition Systems Using Neural Network Technologies*, I. Guyon and P. S. P. Wang (Eds). World Scientific.

McClelland, J. L. (1976) Preliminary letter identification in the perception of words and nonwords. *Jour. of Exp. Psych.: Human Perc. & Perf., 2*, 80-91.

McClelland, J. & Rumelhart, D.(1981) An interactive activation model of context effects in letter perception: Pt. 1 *Psych. Rev., 88*, 375-

Morton, J. (1969) Interaction of information in word recognition. *Psychological Review, 76*, 165-178.

O'Regan, J. & Jacobs, A.(1992) Optimal viewing position effect in word recognition. *Jour. of Exp. Psych.: Human Perc.& Perf., 18*, 185-197.

Rayner, K. (1986) Eye movements and the perceptual span in beginning and skilled readers. *Jour. of Exp. Child Psych., 41*, 211-236.

Rayner, K. (1979) Eye guidance in reading. *Perception, 8*, 21-30.

Rayner, K., Murphy, L., Henderson, J. & Pollatsek, A. (1989) Selective attentional dyslexia. *Cognitive Neuropsych., 6*, 357-378.

Rayner, K. & Pollatsek, A. (1989) *The Psychology of reading.* Prentice Hall

Reicher, G. (1969) Perceptual recognition as a function of meaningfulness of stimulus material. *Jour. of Exp. Psych., 81*, 274-280.

Rumelhart, D., Hinton, G., & Williams, R. (1986) Learning internal representations by error propagation. In D. Rumelhart and J. McClelland, *Parallel. Distributed Processing, 1.* MIT Press.

Solomon, R. & Postman, L. (1952) Frequency of usage as a determinant of recognition thresholds for words. *Jour. of Exp. Psych., 43*, 195-210.