USING READING MODELS FOR CURSIVE SCRIPT RECOGNITION

M. Côté [1-3], E. Lecolinet [1] , M. Cheriet [2], and C.Y. Suen [3]

[1]Télécom Paris, Département signal, 46 rue Barrault, 75634
Paris Cedex 13, France
[2] E.T.S, Université du Québec, 4750 Avenue Henri-Julien,
Montréal, Qué., H2T 2C8, Canada
[3] CENPARMI, Concordia University, Suite GM-606,
1455 de Maisonneuve West, Mtl, Qué., H3G 1M8, Canada
e-mail: cote@cenparmi.concordia.ca, elc@sig.enst.fr,
cheriet@gpa.etsmtl.ca,
suen@cenparmi.concordia.ca

Abstract

This paper presents a new perception-based model for reading cursive script. We limit the scope of our study to the off-line recognition of isolated cursive words. Starting from observations and assumptions used in the elaboration of reading models, we describe the organization of our pseudo-neuronal system and show the role of the activation mechanism in perceiving and reading cursive script. We have introduced into our model some characteristics specific to cursive script. First, we use appropriate features such as ascenders, descenders and loops. Second, we deal with the ambiguity of letter location by introducing the concept of fuzzy position. The location as well as the missing letters are deduced from the context (i.e. the word-letter lexicon). After implementation of our method, preliminary qualitative results have been obtained and are discussed. We are concentrating now on further formalizing and generalizing the proposed model on a larger data base.

1 Introduction

Teaching the computer how to read is not an easy task especially when the knowledge to train the machine is still not well identified. Since humans are able to read handwritten text with apparent ease, it seems appropriate to base an automatic handwriting reader on human reading models. What features are detected while reading? How do we access information concerning the meaning of a word? Does perception of a word build up from the perception of its letters or by their shapes? For many years, researchers in the fields of biology, neurophysiology, cognitive psychology and linguistics, have studied these questions [1] and reading models have resulted from their investigations. Even if these models are still in progress and many theories defending different ideas are being debated, we believe we can benefit from their observations for the off-line recognition of isolated cursive words. In fact, we consider that features detected while reading, play a key role in determining the choice of good features to look for when performing the recognition. Also, we assume that the problem of mental lexical access can influence the architecture that will be selected for the method to be developed. Though several models have been proposed to explain the process of mental lexicon access while reading [2], such as the logogen model [3], the verification model [4] and the interactive-activation model [5], they mostly rely on printed text. Few studies have been conducted on the mechanisms involved in the reading of handwriting. In [6], the authors conclude that even if the reading of handwritten words differs from printed words at first glance, once cursive normalization has taken place, handwritten and printed words seem to be subject to simular processes. For example, experimental studies suggest that people use word shape to help them recognize words and this can be useful too in handwriting recognition since the overall shape of a handwritten word follows the same sequence of tall (ascenders: "l", "t"), short ("o", "r", "a") and projecting letters (descenders: "p", "j") than printed word. Thus, it seems that the underlying concepts of reading models can be adapted to the recognition of cursive script.

In this perspective, on-line cursive script recognition has been studied in [7] using a blackboard architecture combining several lexical access models. Off-line cursive script recognition was performed in [8], using a top-down approach, and in [9], using a hypothesis generation, hypothesis testing and global contextual analysis, to name a few. In our method, we select one of the lexical access model used in [7]. We also combine concepts evocated in [8] and [9]: top-down process, hypothesis generation, hypothesis testing and contextual analysis.

Dealing with off-line cursive script does not simplify this adaptation. Because cursive writing is fundamentally ambiguous, off-line cursive script recognition constitutes a complex problem. Often, letters in the words are poorly written if not missing. Consequently, perfect letter segmentation is impossible and letter position is not known precisely.

This paper presents a new method based on a human reading model, influenced by the work of McClelland & Rumelhart in perception. Our model [10] shares with McClelland & Rumelhart [5] the following characteristics: pseudo-neuronal network, parallel processing, and activation mechanism (top-down and bottom-up processes). We will show how we merge perception concepts into handwritten word recognition.

The organization of the paper is as follows: section 2 introduces McClelland & Rumelhart's perception model, section 3 presents our proposed method, section 4 defines activation concepts, section 5 describes the labelling technique in the bottom-up process; results from simulations are shown in section 6. The last section concludes the presentation.

2 Perception model

In their research in psychology, McClelland & Rumelhart developed a model of perception. Their assumptions were twofold: (1) perceptual processing takes place within a system in which there are several levels of processing, each concerned with forming a representation of input at a different level of abstraction, and (2) visual perception involves parallel processing. Their model is a neuron-like system based on three hierarchical levels (features, letters and words) with feedback between them. Communication between levels is excitatory and inhibitory. They used this model to find one missing letter in a four-letter word. The words consisted of printed letters described with very simple features (horizontal, vertical, and diagonal strokes). The main psychological justifications of this architecture are [11]:

1. Gradual propagation of activation

Propagation of activation is not all-or-none but rather builds up gradually over time. Their assumption is that both the stimulus and the context accumulate gradually over time. So when both of them are accumulating at the same time, a response threshold is reached more quickly in perception of a letter in a word compared to the perception of a letter in isolation.

2. Activation function

The activation of a unit or a cell (associated with a feature, a letter or a word) is represented by a sigmoid function as shown in Fig. 1, because of its similitude with the behavior of biological systems [1].

[1] "it is often observed in biological systems that tow stimuli presented separately to the system provoke no response, although when presented simultaneously, a response is obtained. Furthermore, once the system is responding, further stimuli are responded to in a linear fashion"[12] p.420 and 421.

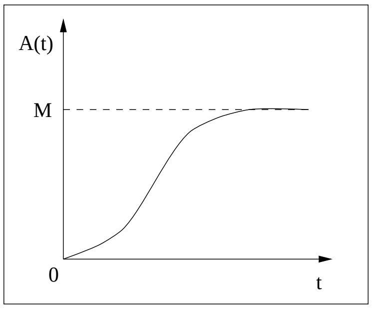

Figure 1: **Sigmoid model of the activation function.**

3. Interactive Processing

In reading, decisions at each level reflect influences from all other levels. So they developed a model that applied this assumption to visual letter recognition. This is why between-level connections are bidirectional and excitatory. Here again, this assumption is motivated by the fact that perception of a letter in a word is facilitated more when it occurs in context than when it occurs in isolation.

4. Mutual competition

Competitive interactions among units allow contrast enhancement and the suppression of weak activations. In their model, McClelland & Rumelhart implemented this idea at each level of processing in order to distribute decision making throughout the information processing system rather than to centralize this function in a limited capacity executive. This is why within-level connections are bidirectional and inhibitory.

3 Proposed method and architecture design

Although McClelland & Rumelhart's architecture is very attractive, it remains limited and cannot be directly applied to cursive script recognition because of the imperfect nature of cursive script compared to the printed texts they were working on. Moreover, as inderlined in [1], their model does not explain among others things, the location uncertainty of features and letters, and the outside-in recognition process. Coping at least with these two perceptual factors, our method proposes an original solution to the cursive script recognition problem.

In order to deal with handwritten words, we have introduced into our model additional characteristics specific to cursive script. First, we use features such as ascenders and descenders instead of McClelland & Rumelhart's simple features. Second, we discard the inhibition mechanism in order to keep all the information available until the very last moment, when a decision on the identity of the unknown word is finally made. Thus, we modify their equations of activation (*cf.* Section 4.2). Third, we deal with the ambiguity of letter location by introducing the concept of fuzzy position. The location as well as the missing letters are deduced from the context (*i.e.* the word-letter lexicon). Fourth, we also estimate the number of letters in the unknown word since we work with words of different lengths (instead of McClelland & Rumelhart's four-letter words).

Our system is based on three levels of detectors, hierarchically organized at feature, letter, and word levels as shown in Fig. 2. We assume that there is a detector for each word in the considered lexicon as well as for each letter and for each feature associated with a given location in the image.

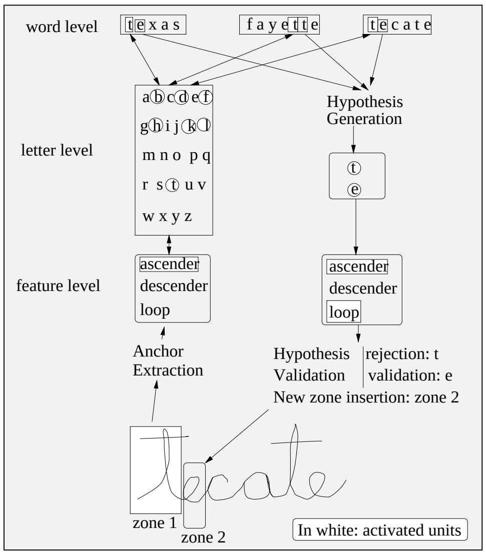

Figure 2: **An example of the activation mechanism within the three-level system for the city name "Tecate".**

Since the letters in the image are in a relative order, we model the letter position by a fuzzy area which does not necessarily contain one letter. We call this area a "zone" which could be envisigned as a window superimposed on an image.

Interaction between detectors consists of excitatory activation of a neural type. Each detector can be seen as a single unit, and the others to which it is connected are called its neighbors. Each connection is two-way.

Connections may occur between adjacent levels only (feature $\longleftrightarrow$ letter, letter $\longleftrightarrow$ word, word $\longleftrightarrow$ letter). Communication consists of excitatory messages which increase the activation level of their receivers. Communication between adjacent levels proceeds through a spreading activation mechanism, in which activation at one level spreads to neighboring levels by following two processes: a bottom-up process (feature level $\longrightarrow$ letter level $\longrightarrow$ word level) and a top-down process (word level $\longrightarrow$ letter level $\longrightarrow$ feature level). Thus, units sharing the same level compete with each other but without any mutual influence.

From an off-line image of an unknown handwritten word (unknown pattern), meaningful features such as ascenders, descenders, and loops are first extracted. These are the anchors of the pattern, like the key letters described by [13]. For each one, a zone is created which is initialized with the corresponding detected features in its area.

In the bottom-up process, letters compatible with the anchors are activated (feature level towards letter level). Up to this point, all the letters of the unknown pattern are not necessarily recognized. Following the same process, activated letters trigger the words they are related to.

When the bottom-up process is completed, the top-down process begins. The activated words generate letter hypotheses which are verified against the features. If the features matching the letter hypotheses are present in the unknown pattern, the hypotheses are validated and the corresponding detectors are activated. Hypothesis generation is therefore the way additional contextual information is brought back to the letter level and then to the feature level. This process helps the segmentation of the unknown pattern; because this segementation is deduced from the relative letter order given by the context. In Figure 2, the hypotheses would be "e" and "t" after zone 1 and nothing before zone 1. The features after zone 1 correspond to the features of the letter "e", and not to the features of the letter "t". So hypothesis "t" is rejected but hypothesis "e" is validated. Zone 2 will then be created and inserted after zone 1. This zone will have letter "e" activated. Thus, the new zone list for the next cycle will be: zone 1, zone 2.

A cycle has been completed when a bottom-up process is followed by a top-down process. The activation of the detectors at the end of a cycle constitutes the input for the next cycle. This input is thus modified (after hypothesis validation) by the contextual information wich triggers the insertion of new zones. After several cycles of top-down and bottom-up processes, the system saturates and the process stops automatically. The words having the highest activation value among candidate words are selected as a recognition result. The output is a list of candidate words sorted by decreasing order of activation.

Two lexicons link the different levels of detectors: a word-letter lexicon and a letter-feature lexicon. There is also a matching array for each reference word. This array traces the degree of match between the reference word letters and their corresponding zones in the input image, until the end of processing is reached.

4 Activation

In this section, we describe McClelland & Rumelhart's equations of activation and some of the perceptive phenomena used in their elaboration. Following this introduction, we demonstrate how McClelland & Rumelhart's equations drive us to new equations more adapted to the recognition of isolated cursive words.

4.1 McClelland & Rumelhart's equations of activation

The activation function is represented by the following equation:

$$A_i(t + \Delta t) = A_i(t) - \Theta(A_i(t) - r_i) + E_i(t) \quad (1)$$

If the contribution of the neighbors is globally **excitatory**,

$$E_i(t) = n_i(t)(M - A_i(t)), \quad n_i(t) > 0$$

If the contribution of the neighbors is globally **inhibitory**,

$$E_i(t) = n_i(t)(A_i(t) - m), \quad n_i(t) < 0$$
$$\text{and} \quad n_i(t) = \sum_j \alpha_{ij} a_j(t) - \sum_j \beta_{ij} a_j(t)$$

with:	$A_i(t + \Delta t)$	activation for unit i at time $(t + \Delta t)$.
	$A_i(t)$	activation for unit i at time (t).
	Θ_i	a decay rate constant for unit i.
	r_i	the resting level of unit i.
	$E_i(t)$	effect on unit i.
	$n_i(t)$	net input on unit i.
	α_{ij}	excitatory connection weight between node j and i.
	β_{ij}	inhibitory connection weight between node j and i.
	$a_j(t)$	activation for unit j, neighbor of unit i at time (t).
	M = 1	maximum activation level.
	m = -0.02	minimum activation level.

Eq. (1) models the process of lexical access in humans [2]. Thus, it takes into account perceptive phenomena encountered during the process of human reading. Amoung other things, it accounts for:

- The word frequency effect
- The word similarity effect
- The word repetition priming effect.

- **The word frequency effect and the parameter r_i**

In general, subjects will recognize more rapidly common words such as "**HOUSE**"than unusual words such as "**HOUND**". McClelland & Rumelhart, use the resting level to model this effect. The resting level of the activation function will then be dependent on the word frequency. A low-frequency word will have a lower resting level. It will then need more stimuli before having enough energy to be activated. On the contrary, a high-frequency word will need less stimuli to be activated, as is illustrated in Fig. 3. Therefore, in the system, the strength of the connection between the letter unit H and the word unit HOUSE will be greater than between the letter unit H and the word unit HOUND.

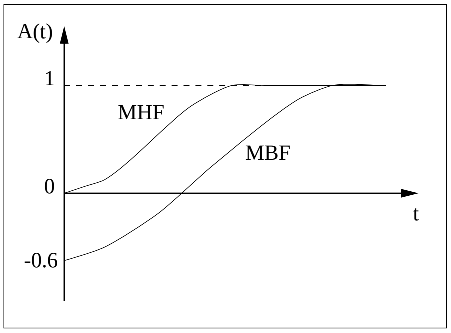

Figure 3: **Resting level and word frequency. MBF is a low-frequency word, $r_i = -0.6$; MHF is a high-frequency word, $r_i = 0$.**

- **The word similarity effect and the inhibitory term, $\sum_j \beta_{ij} a_j(t)$**

Each word in the lexicon is part of a subset of words showing similarities. For example, the word $word \in \{word,\ work,\ sword,\}$.

The use of inhibition permits sharpening the response coming from the system. The increase of the activation of a unit associated with a word will inhibit the units associated with inappropriated words.

- **The word repetition priming effect and parameter Θ**

When a word has been recognized, it will be more rapidly identified subsequently. McClelland & Rumelhart, use the constant Θ, the activation decay rate, to model this effect. When a word has already been activated, its activation decreases slowly. Hence, when it is elicited again, fewer stimuli will be necessary to increase its activation.

4.2 New equation of activation

Three hypotheses have been established in order to adapt this equation to our needs:

H1: same word frequency for all the words in the lexicon

For now, we do not take into account word frequency.

H2: no inhibition

We do not take into account the word similarity effect. In the cursive script context, information is often missing. Thus, we keep all the information available until the very last moment, when a decision is made on the identity of the unknown word. Possible solutions should not be eliminated too early in the decision process.

H3: decay rate for the activation

We retain McClelland & Rumelhart's decay rate of activation, which allows the activation fuction to reach saturation.
Hypotheses H1, H2 and H3 drive us respectively to the equations of Table 1.

$r_i = 0$
$\sum_j \beta_{ij} a_j(t) = 0$
$\Theta = 0.07$

Table 1: **Hypotheses supporting our activation equation.**

In order to show the originality of the equations used in our method, Table 2 compares McClelland & Rumelhart's equations with ours.

If $n_i(t) > 0$, $A_i(t+\Delta t) = (1-\Theta)A_i(t) + \Theta r_i + n_i(t)(M - A_i(t))$ If $n_i(t) < 0$, $A_i(t+\Delta t) = (1-\Theta)A_i(t) + \Theta r_i + n_i(t)(A_i(t) - m)$ with $n_i(t) = \sum_j \alpha_{ij} a_j - \sum_j \beta_{ij} a_j(t)$
$n_i(t) > 0$ in all cases, and $A_i(t+\Delta t) = (1-\Theta)A_i(t) + n_i(t)(M - A_i(t))$ We set $n_i(t) = \mathcal{E}_i(t) = \sum_j \alpha_{ij} a_j \Rightarrow$ excitatory contribution

Table 2: **On top, McClelland & Rumelhart's equations; below, ours. Note that α_{ij} is defined differently in the two cases.**

In eq.(1), for convenience we set $n_i(t)$ as the excitatory contribution, $\mathcal{E}_i(t) = \sum_j \alpha_{ij} a_j$, and we obtain,

$$A_i(t+\Delta t) = (1-\Theta)A_i(t) + \mathcal{E}_i(t)(M - A_i(t)) \tag{2}$$

A unit is active if its activation is positive. Inactive units have no influence on their neighbors. In the absence of inputs from their neighbors, all units are assumed to decay back to a resting level (activation = 0). Each unit has a momentary activation $A_i(t)$. The unit's activation varies between 0 and 1 in accordance with the following equation:

$$A_i(t+\Delta t) = A_i(t) - \Theta_i A_i(t) + E_i(t) \tag{3}$$

where:		
	$A_i(t+\Delta t)$	the new value of the activation of a unit i at time $(t+\Delta t)$.
	Θ_i	a constant for the unit's decay.
	$E_i(t)$	the effect on unit i at time t due to inputs from its neighbors.

The effect from the neighbors on a unit i at time t, $E_i(t)$, is represented by:

$$E_i(t) = n_i(t)(M - A_i(t)) \tag{4}$$

with: $n_i(t)$ the excitatory influences from the neighbors at time t on unit i.

M the maximum activation level of the unit, set to 1.

where the factor $n_i(t)$ is defined as:

$$n_i(t) = \sum_i \alpha_{ij} a_j(t) \tag{5}$$

with: $a_j(t)$ the activation of an active excitatory neighbor of the unit.

α_{ij} the associated weight constant.

The factor $(M - A_i(t))$ modulates the contribution of the neighbors $n_i(t)$, to keep the input to the unit from driving it beyond some maximum. As can be seen, when the activation of the unit has reached its maximum value (one), the effect of the input is reduced to zero. Thus, the activation is always bounded. Then, with eq. (4) we can rewrite eq. (3) as:

$$A_i(t + \Delta t) = (1 - \Theta)A_i(t) + n_i(t)(M - A_i(t)) \tag{6}$$

As can be demonstrated, eq. (6) will saturate if Θ is positive and < 1. We set $\Theta = 0.07$ as in [5] because this value gives a reasonable time constant for saturation.

Our system is defined as pseudo-neuronal because the weights are not learned. They adapt during the process according to the following formulas:

$$\alpha_{fl} = \tfrac{1}{NF}, \;\; \alpha_{lw} = \mathcal{F}(\Delta)_{lw} \, \tfrac{1}{NZ}, \;\; \alpha_{wl} = \tfrac{1}{NW}$$

where NF is the number of features found in the signal for letter L, $\mathcal{F}(\Delta)_{lw}$ is the position coefficient for letter L in word W (*cf.* Section 5), NZ is the number of zones found in the signal at time t, and NW is the number of words in the lexicon containing letter L.

5 Labelling technique

During the bottom-up process, we first try to match the activated letter with the related word (letter-word lexicon), and at the right position within the word. To do so, a labelling technique has been developed which deals with the relative order of letters and the letter's fuzzy position in a word.

Letter labelling is done starting from the ends of the word moving in, since this has been shown to be the way humans do it during the early stages of reading [1]. Letters described with meaningful features and with high activation are first labelled.

The estimation of the relative letter positions is a very important parameter for a successful letter labelling of each word in the lexicon. During the labelling process, we always compare the position of the tested letter in the word with the corresponding position in the image. Since the letter position in the image is not known precisely, we introduce a position coefficient (real number between 0 and 1). When its value is 1, we consider that the letter in the word corresponds exactly with the letter in the image. On the contrary, a 0 value means that the letter in the word does not correspond at all with the letter in the image. Between these two limits, this coefficient varies in accordance with function $\mathcal{F}(\Delta)$ shown in Fig. 4.

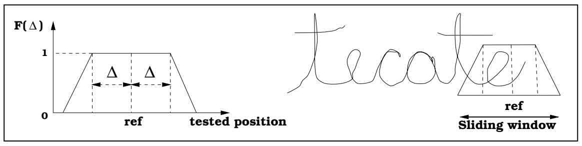

Figure 4: **Position coefficient function.**

Here, the reference is the letter position within the zone, Δ corresponds to the spreading of the letter in the image (word length divided by the estimated number of letters in the unkown pattern), and the tested position is the projection of a word's letter on the image. One can picture this function as an observation window of width 4Δ sliding across the image.

6 Simulation and preliminary results

At the present time, feature extraction is performed using a method which provides the input of our system with a sequence of features and their location in the image. In the method used in [8], the feature lexicon includes 27 different features. Based on observations, these features have been organized into two categories: the very meaningful and the others. There are 4 features of the first category: ascenders, descenders, loops and e_strokes. The latter category includes unstable loops in "d, p, b, q", and strokes in "m, n, w". Because those features are usually missing, their presence is attested only when they are found in combination with the first class of features. For this reason, they are called conditional features. Feature identification is thus performed in two steps: find the meaningful features, and identify the conditional features in relation with the meaningful ones.

The word lexicon contains 60 words with 3 to 9 letters. The results for words "Tecate", "Ithaca", "Ten" are shown in Fig. 5. In the second image, the segments are the virtual representation for strokes and ascenders. The loops are represented by circles. In the third image, each zone is represented by a box. As we can observe, the meaningful features such as ascenders and loops have been detected, and are associated with good conditional features. Moreover, the word "Tecate" is at the top of the ouput list.

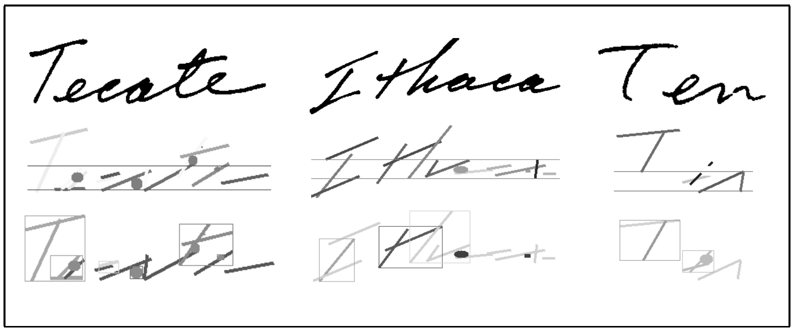

Figure 5: **Words "Tecate", "Ithaca", and "Ten" before processing, after feature extraction, and with zones created after the bottom-up and top-down processes.**

With the activation curves shown in Fig. 6, we can follow the behavior of the most activated units (not all the units are represented here) involved in the recognition process of the word "tecate".

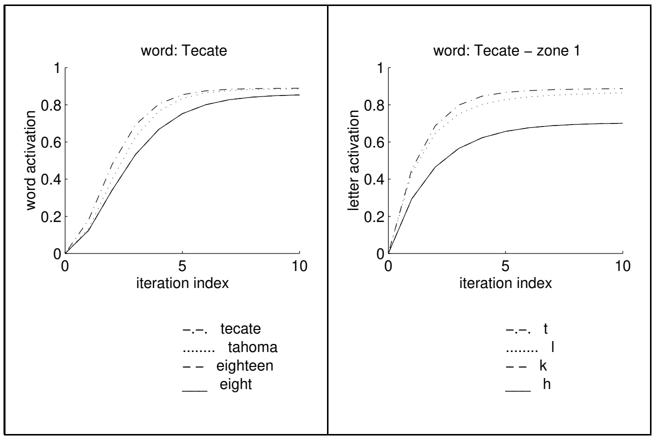

Figure 6: **On the left, activation word curves; on the right, activation letter curves for zone 1, area related to the beginning of the word.**

In the first graph, candidate words, “tecate” and “tahoma”, beginning with the letter “t” are more activated. Both words “eighteen” and “eight” contain the letter “t” but not at the proper location consequently, their activation is less important. In the second graph, letter “t”, “l”, “k” and “h” in zone 1, compete together because of their commun feature. Since fewer words in the lexicon are beginning with the letters “l”, “k” and “h”, the letter “t” gets more activation. This figure illustrates that when a unit is activated and receive feedback from its neighbors of adjacent levels, its activation increases following eq. (3) until saturation is reached.

7 Conclusions

We propose in this paper a perception-based model for reading cursive script. Starting form observations and assumptions used in the elaboration of reading models, we discuss how McClelland & Rumelhart’s lexical access model can be adapted to the off-line recognition of isolated cursive words. We present the organization of our pseudo-neuronal system and show the role of the activation mechanism in perceiving and reading cursive script. We also incorporated into our model three elements directly suggested by research in perception: significant features in word recognition such as ascender and descender, the fuzzy position concept to deal with the location uncertainty of features and letters, and outside-in labelling proces which mimics the outside-in recognition process observe in experiments with subjects. Our recognition method has been implemented and tested with simulated data. At the moment, we are testing this method on real images from the Cedar database []. In this perspective, we are developing an original extraction module, more adapted to our method and based on Cheriet & Suen’s “key letters”.

Acknowledgements

We would like to thank Claudie Faure and Christine Nadal for their assistance, and financial funding from the IRIS National Networks of Centres of Excellence of Canada, and Natural Science and Engineering Research Council of Canada.

References

[1] I. Taylor and M. M. Taylor, editors. *The Psychology of Reading*. Academic Press, p. 207, 1983.

[2] M. Taft, editor. *Reading and the Mental Lexicon*. Lawrence Earlbaum, Hillsdale, New Jersey, 1991.

[3] J. Morton. Interaction of information in word recognition. *Psychological Review*, 76:165–178, 1969.

[4] C. A. Becker. Allocation of attention during visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 2:556–566, 1976.

[5] J. L. McClelland and D. E. Rumelhart. An interactive activation model of context effects in letter perception. *Psychological Review*, 88:375–407, 1981.

[6] C.M. De Zuniga, G.W. Humphreys, and L.J. Evett. Additive and interactive effects of repetition, degradation, and word frequency in reading of handwriting. In D. Besner and G.W. Humphreys, editors, *Basic processes in reading: Visual word recognition*, pages 10–33. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, 1991.

[7] C. A. Higgins and P. E. Bramall. Handwriting recognition based on the human reading process. In C. Faure, P. Keuss, G. Lorette, and A. Vinter, editors, *Advances in Handwriting and Drawing: A Multidisciplinary Approach*, pages 155–167. Europia, Paris, 1994.

[8] E. Lecolinet. Cursive script recognition by backward matching. In C. Faure, P. Keuss, G. Lorette, and A. Vinter, editors, *Advances in Handwriting and Drawing: A Multidisciplinary Approach*, pages 89–91. Europia, Paris, 1994.

[9] J.J. Hull. A computational theory of visual word recognition, Ph.D. dissertation. Technical report, Department of Computer Science, State University of New York at Buffalo, 1988.

[10] M. Côté, E. Lecolinet, M. Cheriet, and C.Y. Suen. Building a perception based model for reading cursive script. In *Proc. of the third ICDAR95, Vol. II*, pages 898–901, Montreal, Canada, August 14-16 1995.

[11] J.L. McClelland. Toward a theory of information processing in graded, random, interactive networks. Technical report, Parallel Distributed Processing and Cognitive Neuroscience, Carnegie Mellon University, 1991.

[12] D.E. Rumelhart and J.L. McClelland, editors. *Parallel Distributed Processing; Explorations in the Microstructure of Cognition, Vol. 1: Fondations*. The MIT Press, Cambridge, 1986.

[13] M. Cheriet and C. Y. Suen. Extraction of key letters for cursive script recognition. *Pattern Recognition Letters*, 11:1009–1017, 1993.