On Handwriting Recognition System Performance: Some Experimental Results

by
Paulo J. Santos, Amy J. Baltzer,
Albert N. Badre, Richard L. Henneman,
and Michael S. Miller

GIT-GVU-92-13
July 1992

Graphics, Visualization & Usability Center

Georgia Institute of Technology
Atlanta GA 30332-0280

ON HANDWRITING RECOGNITION SYSTEM PERFORMANCE: SOME EXPERIMENTAL RESULTS

Paulo J. Santos
Georgia Institute of Technology
College of Computing
Atlanta, GA 30332-0280

Amy J. Baltzer*
NCR Corporation
500 Tech Parkway
Atlanta, GA 30313-2446

Albert N. Badre
Georgia Institute of Technology
College of Computing
Atlanta, GA 30332-0280

Richard L. Henneman
NCR Corporation
500 Tech Parkway
Atlanta, GA 30313-2446

Michael S. Miller
NCR Corporation
500 Tech Parkway
Atlanta, GA 30313-2446

Performance of a rule-based handwriting recognition system is considered. Performance limits of such systems are defined by the robustness of the character templates and the ability of the system to segment characters. Published performance figures, however, are typically based on pre-segmented characters. Six experiments are reported (using a total of 128 subjects) that tested a state-of-the-art recognition system under more realistic conditions. Variables investigated include display format (grid, lined, and blank), surface texture, feedback (location and time delay), amount of training, practice, and effects of use over an extended period. Results indicated that novice users writing on a lined display (the most preferred format) averaged 57% recognition performance. By giving subjects continuous feedback of results, training, and after about 10 minutes of use, the system averaged 90.6% character recognition. Following three hours of interrupted use and with performance incentives, subjects achieved an average 96.8% accuracy with the system. Future work should focus on improving the ability of the recognition algorithm to segment characters and on developing non-obtrusive interaction techniques to train users, to provide feedback and to correct mis-recognized characters.

INTRODUCTION

Rapidly maturing computer-based handwriting recognition technologies have contributed to the increased availability and popularity of pen-based portable computers whose primary mechanism for text input is through handwritten characters. Pen-based gestural interfaces have been shown to be advantageous for several applications and environments [e.g., Sibert (1987), Wolf (1988)]. Handwriting recognition is an important factor in the successful implementation of these systems. The value of a handwriting recognition system is dependent on the degree to which the system can accurately interpret handwritten characters.

Handwriting recognition system vendors typically report their systems can achieve between 90 and 95 percent character recognition. These accuracy reports are based on the recognition of characters that have been collected under very controlled conditions. Typically subjects will be asked to print specific characters when prompted — slowly, one at a time, and in boxes. Because in more realistic settings users will not always leave sufficient space between characters and actual usage conditions may increase the variability of character formation, handwriting recognition performance in natural environments may be considerably less than reported figures. Few data exist with respect to actual situations. A joint project between the NCR Human Interface Technology Center and the Georgia Institute of Technology has resulted in a series of six experiments that collected handwriting data that more closely resembled handwritten characters found in natural environments. The experiments explored interface characteristics that were expected to affect performance of handwriting recognition systems.

Since current handwriting recognition systems are not able to recognize all styles of human handwriting, and since users of handwriting devices may have writing styles that are difficult to modify, one may wonder what can be done to help users increase the recognition accuracy. The purpose of this paper is to identify and explore interface characteristics that may improve recognition system performance and user satisfaction. Interface characteristics investigated in this paper are the surface texture of the writing device, display format, recognized-character feedback, training and practice.

METHOD

Subjects

A total of 128 subjects participated in this series of experiments. Subjects were either Georgia Tech students or NCR employees with an average age of 23.7 years. The population used for these experiments was more homogeneous than the population of potential users of a handwriting-based product. The relative impact of the various independent variables, however, should extend to the actual population.

Apparatus

Subjects wrote on a Seiko D-Scan digitizer/display device. The display was passive matrix VGA covered by a transparent digitizer and a glass plate, and was connected to an NCR 386 PC. A stylus that was tethered to the digitizer was used as the writing device.

* Currently at University of Colorado, Department of Psychology, Boulder, CO 80309.

Task

Each session of the six experiments consisted of variations of the following task. For each handwriting session, subjects copied a paragraph onto a transparent digitizer that rested above an LCD. As the subjects wrote, electronic "ink" appeared on the display directly under their pen. The paragraph they were to copy was projected onto a wall with an overhead projector to mimic a note-taking task. In Experiments 1 through 5, the same paragraph was used for all subjects; in Experiment 6, in which subjects visited the laboratory for 14 sessions, a different paragraph was used for each session.

Except for Experiments 1 and 2, subjects were given brief instructions on how to form characters such that they would be recognized by the system. Depending on the amount of training they were to receive, subjects were also shown a guideline sheet for forming characters and were allowed to refer to this sheet when writing. Each subject was instructed to write in all block, uppercase letters (the only type of input recognizable by the system). System software recorded the location, order, and timing of each handwritten stroke.

Procedure

All subjects were told to write in block, uppercase letters (the only type of input recognizable by the system). To create a more realistic task scenario, subjects were given more instructions on how to form characters after the first two experiments. They were told to letter neatly and to leave a small space between each character and at least a 1/2 inch space between words. For the final three experiments, subjects were also shown a guideline sheet that had examples of acceptable character formations for each of the letters in the alphabet. Subjects were allowed to refer to this sheet when writing.

Table 1 provides a summary of the design of each of the six experiments. The following independent variables were manipulated in one or more of the six experiments reported:

Display format: The most common type of paper used for writing is lined paper. Other types used are blank sheets (without lines), and sheets of paper with grids. A common type of display used in handwriting recognition systems is a display containing a grid that forces users to separate characters, thereby aiding the segmentation process. However, work by Barnard (1976) has suggested that speed of performance and user satisfaction are best with less constraining formats, such as a lined or blank display. It was expected that the type of display used may influence the handwriting and satisfaction of subjects. Experiment 1 explored the relative merits of three levels of this variable (grid, lined, and blank.)

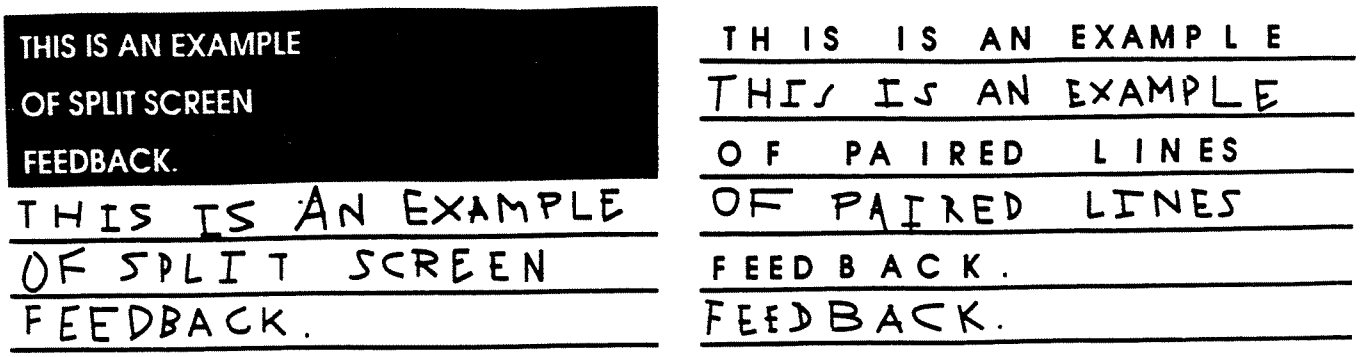
Surface texture: The surface texture of a digitizing tablet can feel quite different to a user than the feel of writing on paper with a pen or pencil. Digitizing tablets are typically made of a hard plastic and are written on with a hard plastic stylus tip. The surface texture, if much different (more slippery or sticky) than pen on paper, could adversely affect human handwriting performance. Experiment 2 measured the effect of three digitizer surfaces at varying levels of glossiness on human performance and preferences.

Experiment	N	Variable	Levels	Performance
1	30	*Display format*	*Grid*	0.76
			Lined	0.57
			Blank	0.60
2	12	Surface texture	High gloss	0.77
			Medium gloss	0.80
			Low gloss	0.81
3	13	*Distant feedback*	*Before feedback*	0.78
			After feedback	0.82
4	48	*Training*	*None*	0.77
			Short	0.88
			Long	0.89
		Feedback	*Postponed*	0.78
			Continuous	0.82
5	15	Feedback	Distant	0.82
			Close	0.83
6	10	Feedback	Postponed	0.94
			Continuous	0.93
		Practice	1st session	0.90
			2nd session	0.93
			3rd session	0.94
			7th session	0.95
			14th session	0.96

Table 1. Summary of handwriting recognition algorithm performance (fraction of characters correctly recognized)
Note 1: Experiment 2 used gridded screen format, Experiments 3-6 used lined screen format
Note 2: **Variables** in italics had a significant ($p<0.05$) effect on performance. **Levels** in italic were significantly different ($p<0.05$) from the other levels of that variable.

Feedback: "Feedback" refers to the display of the system's interpretation of the user's handwritten characters. Both the location and timing of the feedback were expected to influence human writing performance. The location of the feedback could be displayed adjacent to (above or below) the location the character was written on the display or in some other more distant location. Close feedback (Figure 1a) should be easier to monitor than distant feedback (Figure 1b.) The close feedback may also distract from the writing task. Similarly, feedback could be displayed as a subject writes each character (continuous feedback) or displayed at some later time (postponed feedback). Again, continuous feedback should be easier to monitor, but could also be distracting to the user. Experiments 3 through 6 explored the effects of various types of feedback on performance.

Training: The amount of training a user receives on using the device was expected to affect user performance. Longer training was expected to improve performance. However, the amount of improvement training provides was unknown. Three levels of training were evaluated in



a. Split screen design for distant feedback

b. Paired lines screen design for close feedback

Figure 1. Two screen designs for different feedback locations

Experiment 4: 1) no training (subjects were only told to letter neatly so that someone else could read what they had written,) 2) short training with feedback (subjects practiced writing the characters A-Z and the digits 0-9 three times, with feedback), and 3) long training with feedback (writing the characters A-Z and the digits 0-9 repeatedly with fewer than three errors before continuing.)

Practice: Finally, it is well known that subjects improve on most tasks with practice. Both the amount of performance improvement subjects were able to achieve with practice and the amount of practice needed to achieve a significant increase in performance were unknown. In Experiment 6, subjects visited the laboratory fourteen times, and in each session they wrote a different paragraph using the system. It was assumed that in each session the experience subjects gained during previous sessions influenced them in their behavior and writing style.

The dependent variables of interest were system recognition performance (measured by the fraction of characters correctly recognized), human preferences, and writing time. Results are discussed in the following section.

RESULTS

System recognition performance

Table 1 contains a summary of system recognition performance results from the six experiments. Variables that resulted in a significant main effect are shown in italic type ($p<0.05$). There were no significant interactions between variables.

In Experiment 6, there was a significant difference ($p<0.05$) in performance between sessions that were approximately ten or more sessions apart. These results show that there was slow but continuous improvement in recognition performance. However, the experiment was not carried on long enough to identify a point in the curve which would indicate the upper limit of performance for an expert user.

Data from Experiment 6 were also classified into four error categories: segmentation errors (system groups strokes to form characters incorrectly), misrecognition of characters

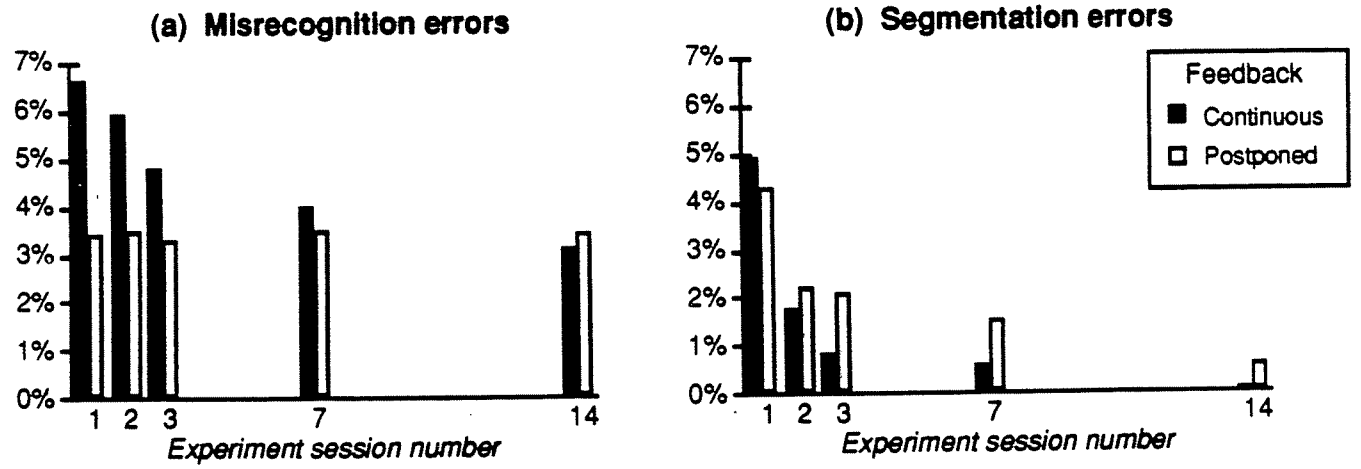


Figure 2. Percentage of characters not recognized due to (a) misrecognition and to (b) incorrect segmentation

(system combines strokes into characters correctly but recognizes characters incorrectly), timeout error (subject pauses for over one second between strokes of a single character), and extraneous characters (subjects inadvertently touches the tablet with the stylus). The first two categories are responsible for over 99% of the errors. Figure 2 shows the number of errors per 100 characters that were attributed to each of the two most important error categories for both the continuous and postponed feedback conditions.

Human preferences

In Experiment 2, subjects were asked to rank their preferences for writing on the three surfaces on a scale of 1 (most preferred) to 3 (least preferred). The mean rankings for low, medium, and high gloss surfaces was 1.46, 2.04, and 2.50, respectively. There was a significant difference ($p < 0.05$) in rankings between the low and high gloss surfaces.

This result was not unexpected, as subjects preferred to write on surfaces that felt most like writing with a pencil on paper. The low gloss surfaces most closely approximated this feel.

Task completion times

Figure 3 contains a summary plot of task completion times in Experiment 6. *Task completion time* is defined as the sum of the times to complete each page of each session. *Time to complete each page* is the elapsed time between the first and last time the stylus touched a single display page. Completion time, therefore, does not include setup time or time between pages. Figure 3 also depicts *writing time* (or pen-down time), which is the total time that the stylus is in contact with the tablet during a session. Writing time does not include pauses in the middle of a page, setup time, or time between pages.

DISCUSSION

System recognition performance

There is a dramatic increase in recognition performance from the lowest accuracy situation (57% in Experiment 1 with lined display) to the highest (96.8% in Experiment 6, continuous feedback, 14th session). Nevertheless, this highest level of *character* recognition implies that approximately only one of every six *words* (of 5 character length) was recognized. Note, however, that 96.8% happens to be exactly the number found by Neisser and Weene (1960) for the pooled best guesses of nine human observers identifying isolated hand-printed characters. This result suggests that further improvements in recognition should be sought using units of handwriting larger than the character.

Four variables significantly improved character recognition performance: gridded display format (Experiment 1), training (Experiment 4), feedback (Experiments 3 and 4), and practice (Experiment 6). When novice users were forced to separate their characters by writing in a grid, recognition performance increased from 60% or less to 76%. It should be noted, however, that the gridded display format was the one least preferred by subjects, a result that concurs with Barnard and Wright (1976). Practice, training and feedback also caused subjects to learn that character spacing is important to recognition performance. The results presented in Figure 2 suggest that the improvement in recognition comes primarily from a reduction of the segmentation errors. Figure 2 also suggests that subjects who were given continuous feedback realized faster that they could improve recognition with better character separation and improve character separation faster than subjects who received continuous feedback.

Experiment 4 data suggest that continuous feedback is better than postponed feedback (78% vs. 82%) for novices. However, Experiment 6 shows that such a difference was not found for experienced subjects. A surprising result of Experiment 6 is illustrated in Figure 2a.

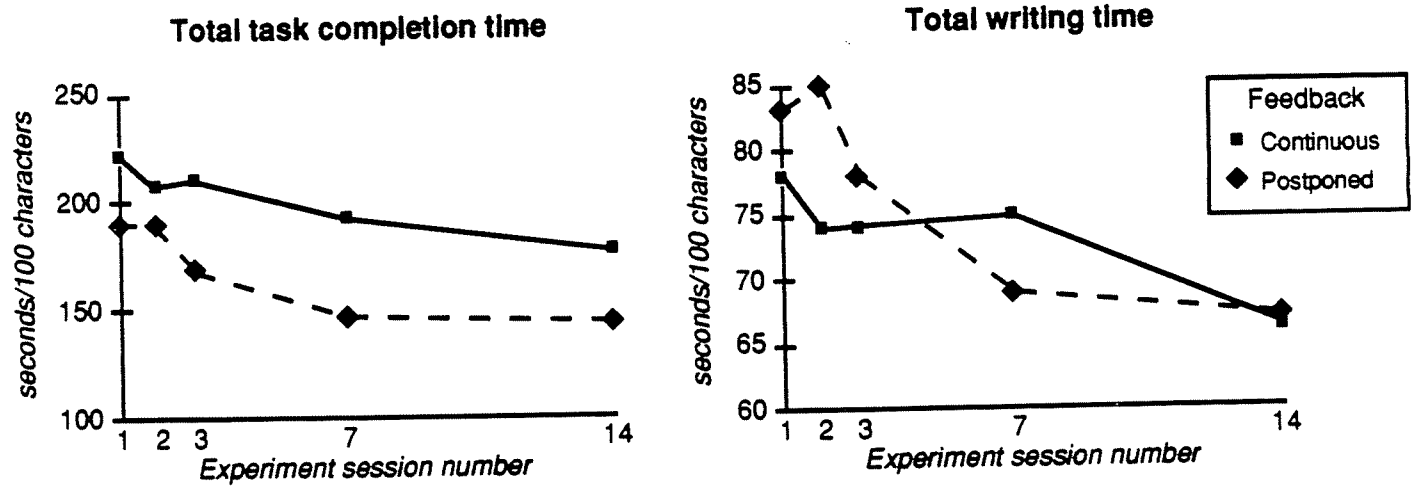


Figure 3. Summary of performance times in Experiment 6 (seconds/100 characters)

Novice subjects who are given continuous feedback form their characters less accurately than subjects who were given postponed feedback, resulting in more recognition errors due to poor character formation. This result could be due to the distracting effect of the feedback. With practice, however, subjects overcame that problem.

These results suggest that effort should be placed on improving the segmentation performance of the handwriting recognition system. Better segmentation is especially useful for first-time and novice users. Poor recognition due to imperfect segmentation may lead to users rejecting the system. Some preliminary effort at segmenting *after* recognition rather than *before* has improved overall recognition performance by as much as 20%. Perfect segmentation before recognition could reduce recognition errors by 45%, sometimes as much as 70%. Approaches such as Fujisaki et al (1991) that take an integrated approach to recognition and segmentation may also significantly improve segmentation and overall recognition accuracy. Timing properties of strokes (e.g., velocity) may also be used to assist in the segmentation process.

Even when segmentation was adequate, however, misrecognition occurred. The results shown in Figure 2b show that users cannot always improve their handwriting: handwriting is, by nature, variable. Thus, the rule base of the system should also be enhanced with additional methods for increasing recognition, such as using character frequency information, contextual information, and dictionary lookup (Burr, 1983.)

Human performance

Caution must be used in justifying the imposition of human performance constraints with data that demonstrates an improvement in character recognition. The user task must determine whether the gain in recognition accuracy is worth the costs of increased writing time, decreased available attention, and decreased user satisfaction.

CONCLUSIONS

In light of the results of these experiments, some general recommendations can be made for applications using handwriting recognition as a means of input. The recognition accuracies reported here appear to be unacceptable for tasks requiring the input of large amounts of data. Unless recognition accuracy can be dramatically improved, handwriting should not be used to replace the keyboard completely. However, for applications where a limited amount of data entry is required, handwritten input may be advantageous.

Constraining users by requiring them to write in all uppercase letters, in grids, on an unfamiliar writing surface, and with character formation rules in mind, may increase the amount of time it takes them to print, and may demand more attention for the writing task, thereby leaving less for other tasks, such as listening and comprehending. Users prefer to write as they normally do, with minimal constraints. Even though constraints will increase recognition, they will also add to user dissatisfaction in using handwriting as a means of input. Therefore, the type of application and environment in which it will be used should determine whether these constraints should be used to improve recognition performance.

Due to current limits in recognition accuracy, valuable data will be lost if initial stroke descriptions are discarded after recognition. Thus, stroke data should be stored so that users may later inspect it. In this way, users may review the handwritten characters themselves, making corrections if necessary.

The system should provide information to the user about how to form characters in a recognizable manner. The approach in these experiments used training and feedback. Training gave users general rules to apply to the formation of all characters, and sample ways to form individual characters. Feedback provided a check for users to see if they were correctly applying the rules they learned in the training. It also allowed the users to try different ways of forming characters on the fly, giving them a means to improve the system's recognition performance. Refinement of these approaches, such as on-line training, should also be investigated.

REFERENCES

Sibert, J. (1987). Issues limiting the acceptance of user interfaces using gesture input and handwriting character recognition (panel discussion), Proceedings of CHI'87, 155-158.

Wolf, C.G. (1988). A comparative study of gestural and keyboard interfaces, Proceedings of the Human Factors Society 32nd Annual Meeting, 273-277.

Barnard, P., and Wright, P. (1976). The effects of spaced character formats on the production and legibility of handwritten names, Ergonomics, 19 (1), 81-92.

Neisser, U., and Weene, P. (1960). A note on human recognition of hand-printed characters, Information and Control, 3, 191-196.

Fujisaki, T., Chefalas, T.E., Kim, J., Tappert, C.C., and Wolf, C.G. (1991). One-line run-on character recognition: design and performance, International Journal of Pattern Recognition and Artificial Intelligence, 5 (1-2), 123-137.

Burr, D.J. (1983). Designing a handwriting reader, IEEE Transactions on Pattern Analysis and Machine Intelligence, 5 (5), 554-559.