Recognition Accuracy and User Acceptance of Pen Interfaces

Clive Frankish*, Richard Hull, Pam Morgan*
Personal Systems Laboratory
HP Laboratories Bristol
HPL-94-84
September, 1994

pen-based input,
handwriting recognition

The accuracy of handwriting recognition is often seen as a key factor in determining the acceptability of hand-held computers that employ a pen for user interaction. We report the results of a study in which the relationship between user satisfaction and recogniser performance was examined in the context of different types of target application. Subjects with no prior experience of pen computing evaluated the appropriateness of the pen interface for performing three different tasks that required translation of handwritten text. The results indicate that the influence of recogniser performance on user satisfaction depends on the task context. These findings are interpreted in terms of the task related costs and benefits associated with handwriting recognition. Further analysis of recognition data showed that accuracy did not improve as subjects became more practiced. However, substantial gains in accuracy could be achieved by selectively adapting the recogniser to deal with a small, user specific subset of characters.

INTRODUCTION

The development of pen interfaces is a key element in strategies for increasing the market for lightweight, hand-held computers. The small size of these devices precludes the use of conventional keyboards, and many are intended to appeal to users who do not have keyboard or computing experience. The preferred technical solution for products such as Personal Digital Assistants (PDAs) is often a graphical user interface in which a pen can be used for pointing and selection functions, drawing, and text entry. In this context, the pen is at least as effective as a mouse for direct manipulation of screen objects. In a well-designed pen interface, screen prompts, user action, and system feedback will be integrated in an immediate and intuitive manner. Direct graphical input, such as freehand drawing, is also much easier with a pen than with a mouse. However, the virtues of handwriting recognition as a means of entering text are less certain. The idea of handwritten input is attractive, particularly for inexperienced computer users, but is offset by the need to correct recognition errors. There are also dissimilarities between human perception and machine recognition which can cause frustration when handwriting recognisers behave in ways the user does not understand.

In assessing the potential for pen based computing, the unreliability of handwriting recognition is the most obvious limiting factor. For system developers, improved accuracy is therefore a major goal. However, it is far from clear what we should set as a realistic target for this effort. There are also very few empirical data to indicate what gains we might expect in user acceptance for a given increase in recognition performance. One aim of the study described here was to provide some quantitative information about the relationship between recognition accuracy and user satisfaction, for different types of pen application. The results confirmed that this relationship is highly task-dependent. In the light of these findings, error data from the test applications were further analysed to assess gains in recognition accuracy that might be achieved as users become more practised, or through limited adaptation of the recogniser to individual users.

QUESTIONS ADDRESSED IN THE STUDY

How important is recognition accuracy in determining user acceptance of pen interfaces?

Few could disagree with the assertion that "the value of a handwriting recognition system is dependent on the degree to which the system can accurately interpret handwritten characters" [2]. This might be taken to imply that high targets should be set for acceptable levels of system performance, and this appears to be supported by data from some evaluation studies in which text copying tasks have been used to assess recogniser performance. In one such study, the general opinion of subjects who had achieved a mean accuracy of 93.2% was that the recogniser would not be of practical significance unless it could be made both faster and more accurate [3].

However, user evaluation of handwriting recognition systems is not necessarily a reliable indication of the acceptability of applications in which these systems are used. In most pen applications, translation of handwritten text will be only one component of user interaction. Users are likely to judge the value of an application primarily in terms of its appropriateness for completing a task, rather than their satisfaction with this one aspect of pen function. The impact of recognition errors will therefore depend upon factors such as the amount of text entry required, and the benefits of using handwritten text input as compared with other available methods of data entry. The importance of recognition accuracy can therefore only be assessed in a broader context that takes into account the nature of the task that users are trying to perform. Within this context, evaluation studies should attempt (a) to establish minimum levels of acceptable performance, and (b) to assess the extent to which further gains in accuracy above this minimum level will result in increased user satisfaction.

To what extent is recognition accuracy under users' control?

The performance of handwriting recognisers is partly determined by individual differences in neatness, consistency, and usage of idiosyncratic letter forms. These attributes will also determine the gains in accuracy that can be achieved by using handwriting samples to adapt the recogniser to individual users. Recogniser 'training' will generally produce greatest benefits for users who produce letter forms that are idiosyncratic, consistent, and mutually distinctive.

For most individuals, handwriting characteristics are well established and stable. Indeed, one of the attractions of pen interfaces is the prospect of using familiar pen and paper skills for interacting with computers. However, handwriting skills do not transfer completely from paper to screen. New users must adjust to differences in the pen and writing surface, and more importantly, to constraints imposed by the recogniser. The amount of learning that occurs, the time it takes, and the effects on recognition accuracy are all important in determining how potential users should be introduced to pen interfaces. Adjustment of motor skills is likely to be fairly rapid. However, deliberate modification of writing styles to improve recognition accuracy is more problematic, even for motivated users. Feedback from the recogniser could in principle allow users to develop an internal model of the recognition process, and to adapt their writing styles accordingly. If

this were happening, recognition accuracy should improve with practice. Available data on practice effects suggest that this improvement is perhaps rather marginal. One reported study found no evidence that experience with a recognition device caused subjects to change their writing styles [3]. In this case no differences were detected between 'ink' generated by subjects during an initial period of pen familiarisation, and that produced after extended practice with the recogniser. In a longer study, the frequency of misrecognition errors remained constant over a sequence of 14 test sessions, although subjects did manage to reduce segmentation errors by improving their control of letter spacing [2]. Finally, there is some evidence that users will be more successful in learning to control their writing styles if some form of explicit instruction or support is provided. learning For example, recognition accuracy might only improve with practice if subjects are allowed to inspect their training prototypes during the test session [5].

In these studies, the amount of explicit instruction and pre-trial practice was variable and not always clearly reported. One aim of the present investigation was to monitor the performance of an untrained recogniser from subjects' first attempts at handwritten input, through an extended period of use. The extent of user adaptation to the system was measured both in terms of overall accuracy and the relative accuracy of first and second attempts at recognition. First attempts may fail because of execution errors which cause characters to be poorly formed, or because users revert to earlier writing styles. However, second attempts at entering misrecognised characters are much more carefully controlled. Users who have developed an accurate model of the recognition process should be able to utilise this model to improve the accuracy of their re-entry attempts. If this were happening, we would expect to find that when all other factors are controlled for, second attempts at data entry are generally more successful than first attempts.

How can recogniser performance be improved?

For recognisers designed to identify discrete, handprinted characters, the problem is one of discriminating between patterns within a limited set of possible alternatives. Because discrimination will generally be easier for smaller set sizes, system designers can improve recognition accuracy by designing interfaces in which input fields will only accept restricted subsets, such as digits or lower case letters. This is an effective strategy for some types of application, particulary those that involve form-filling tasks where the format of information is well defined.

With unrestricted input of handprinted characters, the performance of the current generation of recognisers approaches the accuracy achieved by humans, which might be taken as a theoretical limit. One much-quoted study reports an accuracy of 96.8% for human identification of

carefully handprinted upper case letters, viewed in isolation [1]. Mean levels of accuracy at around these levels have been reported in laboratory tests, using copying tasks [2], [4]. However, these very high levels of accuracy tend to be obtained for trained recognisers, often with large training sets that have been elicited from users under supervision. For devices intended for inexpert users, there is inevitably a question whether lengthy procedures for recogniser training will be fully or appropriately completed. In the present study, untrained recognition data from individual subjects was examined in order to assess the gains in accuracy that might be achieved by a selective retraining procedure targeted on a small subset of poorly recognised characters.

THE EXPERIMENTAL STUDY

For the purposes of this study we devised three test applications, representing different types of task that might be accomplished by means of a pen-based system. Each was implemented on an IBM-compatible PC, using the Microsoft Windows for Pen software environment, with a Wacom PL-100V pen tablet. For the section of the study from which the present data are taken, the recognisers were set to operate in 'boxed, discrete' mode; i.e. all recognised input was written as discrete characters, with the positions of individual characters defined by 'comb' guidelines. The data reported here were obtained from a total of 24 subjects, each using one of three available recognisers. Since differences in the performance of these recognisers were marginal, and are not germane to the results reported here, no distinction will be made between data obtained from different recognisers.

The three test applications were designed to contrast in the following ways:

- the amount of handwriting recognition required for successful task completion,

- error tolerance; i.e.the extent to which the task demanded that all handwritten text be correctly recognised

- the balance between pen use for input of recognised text and other pen functions, such as pointing, menu selection, and creation of non-recognised 'ink'.

The three applications were:

Fax/memo In this application there were three input fields on the display. Two were used for entering recognised handprinted text; one for the name of the message recipient, the other for a six-digit telephone number. Subjects were required to correct any recognition errors within these fields. A third and larger 'scribble' field was used for writing the message itself. Pen traces appearing in

this area were not passed to the recogniser. Completed messages were despatched by a pen tap on a 'send' screen button.

Database This task was organised around a database containing approximately 1500 simulated student records, indexed by name. Access to a particular record was achieved by entering handprinted text in surname and first name fields. The recognised text was matched against the database and the corresponding index region displayed in a scrollable window. Best-fit matching meant that successful access could be achieved by incomplete or partially misrecognised input. Final selection and display of the desired record was achieved by using the pen in point-and-click mode, after which new data (three two-digit examination scores) were entered into the record, again as recognised handprinted text.

Diary This was a standard type of diary application, with a month overview that could be expanded using point-and-click responses to open the appropriate appointments page. Subjects were given a series of brief scenarios, from which they created diary entries in their own style. These entries were entered as recognised handprinted text, and subjects were instructed to correct any recognition errors before closing the diary page.

The three experimental tasks were entirely accomplished by means of the pen interface, and in all three cases the pen was used for point-and-click selection, as well as for input of recognised text. For the fax/memo application, these pen functions were embedded in a task context that focused on the creation of an unrecognised ink trace. In the database task, the combination of point-and-click and recognition of handwritten input provided an economical and effective means of accessing records. The requirement for completely correct recognition was confined to entry of a limited amount of new data. Finally, in the diary task the main focus was on completely correct entry of short text notes (typically around five or six words). In addition to these three simulated applications, the test session also included two sessions of a copying task ('Pen Practice'). This was based on entry of a set of single words or five-digit strings which included the complete set of letters and digits, and was used to obtain a controlled measure of recognition accuracy.

In the experimental session, subjects with no previous experience of pen computing were first shown the pen and tablet, and given a brief description of the principles of pen interaction and handwriting recognition. They then immediately filled in a questionnaire which dealt with their expectations of this type of pen based system. They then had an average of one and a half hours experience with the system. This began with a brief instruction in the use of simple editing procedures, which were practised without the need for character entry. This was followed by

the Pen Practice task, and then sessions with the each of the three test applications, each involving completion of a series of predetermined tasks. The order of these sessions was counterbalanced across subjects. Subjects then completed an evaluation questionnaire which sampled various aspects of attitudes to the pen interface. Finally, recognition accuracy was again assessed using a modified version of the initial Pen Practice task. In the whole course of the experimental session, subjects entered an average of 1023 characters to the recogniser.

DATA

Data logged from the pen tasks included ink traces, all user control actions, and recogniser output. Because subjects occasionally performed incorrect actions (e.g. mis-spellings) scoring of recognition accuracy was based on a transcribed record of the session, in which user inputs were verified by inspection of the corresponding ink trace. The following analysis directly addresses the questions about users' attitudes to recognition accuracy raised in the opening discussion.

How important is recognition accuracy in determining user acceptance of pen interfaces?

Mean recognition accuracy was 87.0% (range 7o 2 - 94.6%). The distribution of individual scores is shown in Figure 1. Further analysis indicated that letter identification was rather better than these figures might suggest. Mean recognition rates for lower and upper case letters were 90.9% and 76.1%, respectively. The relatively high error rate for upper case letters was largely accounted for by case errors, with frequent lower case substitutions for letters such as C, O, S, V, etc., which have identical lower and upper case forms. When operating in 'boxed' mode, recogniser assignment of case is based on the size of the character relative to the comb guide. When case errors were excluded, mean recognition accuracy for upper case letters was 89.4%.

Figure 1: Mean recogniser performance (24 subjects)
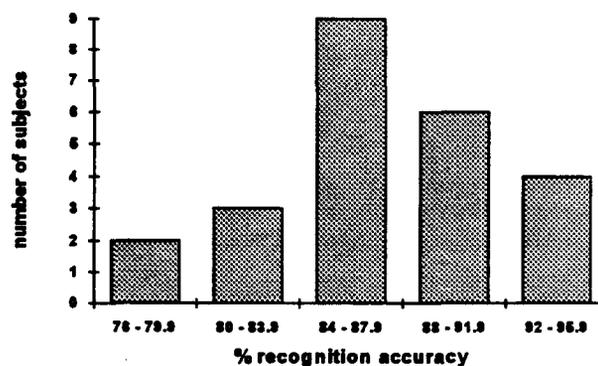
**Table 1:** General appraisal and task appropriateness items from the subjective evaluation questionnaire.

Questionnaire item		mean rating	correlation with recog.
General appraisal			
1. How much do you like the idea of using a pen as a means of working with computers?		6.2	0.21
2. Would you use this type of computer again?		6.0	0.15
3. Did you enjoy using the pen based computer?		6.4	0.33
4. Did you find the pen system you have just used:	frustrating vs. satisfying	4.3	0.41
5.	difficult to use vs. easy to use	5.8	0.23
6.	difficult to understand vs. easy to understand	8.8	0.17
7.	time consuming vs. time saving	4.1	0.35
Mean rating, items 1 - 7		5.7	0.43
Task appropriateness			
How appropriate was the pen-based system for completing the tasks?			
8. Appropriate for fax		6.6	0.36
9. Appropriate for records		8.9	0.22
10. Appropriate for diary		4.8	0.45
Mean rating, items 8-10		6.7	0.48

Average ratings across all 24 subjects suggest a mixed attitude to the pen interface. A high rating for ease of understanding confirms the intuitiveness of the pen interface. Several of the remaining attitude responses were neutral or marginally positive. However, consistent ratings of the pen applications as frustrating and time consuming suggest that recognition errors were perceived as a significant problem. If this is generally the case, we would expect that subjects who managed to achieve high overall recognition accuracy would be more satisfied than those who were less successful. But although the correlations between recognition performance and subjective appraisal are all positive, the relationship is not a strong one. The only correlations with overall accuracy that are statistically significant are for ratings of frustration/satisfaction, and for scores averaged over all seven appraisal items (critical value for $r(23) = 0.41, p<.05$).

The second set of questionnaire items shown in Table 1 asked subjects to rate the appropriateness of the pen interface for completing each of the three experimental tasks. The records application received a very high rating on this scale, as compared with the mildly favourable assessment of the fax application, and low evaluation of the diary. To what extent are these ratings determined by recogniser performance? This question can be answered by examining the correlations between ratings of task appropriateness and recognition accuracy. These show a positive, but relatively weak relationship between the two measures, and the correlation is statistically significant only for the diary task. In Figure 2, mean appropriateness ratings are plotted for each of the three test applications as a function

Seven questionnaire items were designed to elicit general attitudes towards the pen interface. Subjects indicated their responses to these items on a ten point scale, which was then scored so that 10 represented the most favourable, and 1 the least favourable response to each item. A mean rating of 5.5 thus represents the midpoint of this scale. The seven questionnaire items are listed in Table 1, together with the mean ratings, and the correlations between subjective rating and mean recognition accuracy.

Figure 2: Task appropriateness and recognition accuracy
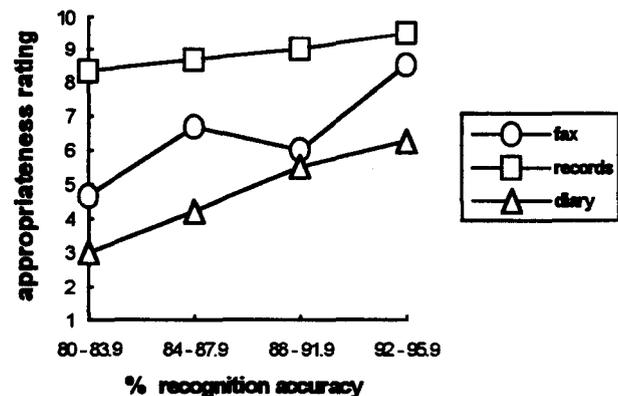
of recognition accuracy. From these data it can be seen that the records application was assessed favourably even by subjects whose handwriting was poorly recognised. In contrast, the diary task was seen as appropriate only by those who achieved high levels of recognition accuracy.

These findings suggest that recogniser performance is a factor in determining the success of pen applications, but that its influence is heavily task-dependent. Some pen applications will be successful despite relatively poor recognition accuracy. The records task used in this study was designed to model a typical information retrieval activity (a target area for PDA applications), where a modest amount of recognised pen input allows the user to achieve significant subgoals within the overall task. In contrast, users may perceive only a marginal advantage in the translation of handwritten diary entries, and this advantage must be weighed against the costs incurred in achieving correct recognition. By reducing these costs to the user, improvements in recogniser performance will be a significant factor in extending the range of potentially successful applications.

How much control do users have over recognition accuracy?

Motivated and compliant users of pen based systems will make considerable efforts to improve recognition accuracy. In the present study, some subjects were clearly experimenting with different writing styles and letter forms. Questionnaire responses reflected a general belief that recognition accuracy did improve with practice. However, the fact that subjects also felt that they had a poor understanding of the causes of recognition failures implies that they did not have an explicit model of the processes involved in machine recognition .

This study was specifically designed to assess the extent to which subjects were able to adapt to the recogniser in the initial stages of practice. The first copying task ('Pen Practice') was completed at the beginning of the experimental session, and was for all subjects their first experience with a recogniser. A matched test was also presented at the end of the session, after subjects had completed a series of trials with each of the test applications. In the event, there were no significant differences in mean recognition accuracy for the first and second Pen Practice sessions (85.3% and 86.4%, respectively; $t(23) = 1.21$).

This finding has two immediate practical implications. First, it suggests that for discrete character recognition, 'out of the box' performance of an untrained recogniser may come close to the standard achievable by a more practised user. The second implication is that unless explicit guidance is given, new users may be unable to modify their writing styles in ways that would allow them to obtain the best possible performance from a particular recognition device.

At another level, the absence of practice effects implies that our subjects were unable to use their experience with the recogniser to develop an internal model of the recognition process. This type of understanding is crucial

in enabling users to respond effectively to input failures. In performing tasks with the pen interface, even the most careful writer will sometimes enter characters that are poorly formed. As with keyboarding errors, these occasional failures need not detract from the fluency of the interaction, provided that the user is able to make an efficient repair. This will be easier if the user understands what was wrong with the first attempt, and knows how to enter a more acceptable form. Thus the accuracy of re-entry may be more sensitive than practice effects as an indicator of the extent to which users possess an accurate model of the recognition process.

In the event, we found no evidence that subjects knew how to modify their writing when recognition had failed. The analysis of first and second entry attempts reported here is for lower case letters only; upper case letters are of less general interest, because the preponderance of case errors is a characteristic of boxed input. To avoid selection bias (letters occurring as second attempts tend to be those that are poorly recognised), accuracy was calculated for each subject by first determining error rates for individual letters, and then taking the average of these values. For mean error rates expressed in this way, there was no evidence of a significant difference in the success of first and second attempts at entry (10.8% and 9.7%, respectively; $t(23) = 1.14$).

Taken together, the lack of improvement with practice and the inability to increase recognition accuracy on re-entry attempts both indicate that subjects' behaviour was not directed by an appropriate model of the machine recognition process. This does not mean that they were incapable of functioning in this way, only that they did not spontaneously use their experience to construct the necessary model. We might therefore conclude that some form of explicit instruction is likely to be necessary if users of pen interfaces are to achieve higher recognition accuracy as they become more practised at using the system.

How can recogniser performance be improved?

There are two main reasons why an untrained recogniser can fail to identify handprinted characters correctly. One is simply the difficulty of discriminating between highly confusable pairs, such as "t" and "f", or "2" and "Z". The other is that individuals may use written forms that are quite discriminable, but do not match the prototypes used to build the recogniser. The best way to deal with the first problem will almost certainly be to introduce 'top down' procedures that use contextual cues to resolve ambiguity. Dictionary lookup is one strategy that is likely to prove effective for many applications. The second problem requires either that users modify their written forms to match recogniser prototypes, or that the recogniser is designed to adapt to individual users. The evidence presented here suggests that users are generally unwilling

or unable to adapt their writing styles to an untrained recogniser. On the other hand, laboratory studies have shown that training the recogniser does significantly improve performance. The disadvantage of this approach is that new users may be required to complete a lengthy training procedure before pen applications can be used.

In order to assess the merits of various approaches to recogniser training, we need more information about costs and benefits to the user. This in turn requires more detailed analysis of the pattern of recognition failures for individual users. For example, lengthy pretraining of the entire character set will not repay the effort involved if either (a) most errors are due to confusable pairs within the recognition set, and discrimination of these pairs does not improve with training, or (b) untrained recognition of most characters is acceptable, but each user has a few idiosyncratic letter forms that are poorly recognised.

The following analysis illustrates the problems that must be addressed, and provides some preliminary answers. The data are based on first attempts at recognition of lower case letters. To reduce sampling errors in estimation of error rates, the data set was further restricted by including only those letters that appeared an average of 12 or more times (range 12.3 - 92.0) in protocols from individual subjects. This excluded b, j, k, q, v, w, x, and z, leaving a total of 18 letters.

For each subject, mean error rates for these 18 letters were calculated, and arranged in rank order. We can then identify the letters that were most problematic for that individual by looking at the top four in this list. Data for the group as a whole give some idea of the extent to which these difficulties are idiosyncratic. If recognition failures were completely due to difficulty in discriminating similar letter forms (e.g. "f" and "t"), then all 24 'top four' lists would be identical. If on the other hand, error patterns were completely idiosyncratic, then the 'top four' lists might look like random selections from the entire set of letters.

The experimental data fall into a pattern that is, not unexpectedly, somewhere between these two extremes. Table 2 shows the number of 'top four' lists that contained each letter, together with the average error rate for the group as a whole. There is clearly some consistency across subjects, and this reflects the difficulty of discriminating letter pairs such as f/t, l/I, h/n, and r/v. But of these, only "f" was a particular problem for a clear majority of subjects. Elsewhere, it is evident that individual subjects had difficulty with letters that were generally quite well recognised. The degree of failure in these cases was often quite extreme; average error rates for the first and second letters in individual ranked lists were 38% and 27%. Overall, these data suggest that for discrete recognition of handprinted characters, there may be advantages in devising a selective procedure for recogniser training,

targeted on a small subset of items that are problematic for the individual user.

Table 2: Letter-by-letter analysis for individual subjects, and mean error rates

letter	'top four' count	mean % error	letter	'top four' count	mean % error
f	18	26.4	y	5	6.2
l	13	12.3	d	3	5.9
h	10	13.9	s	3	5.5
r	10	12.5	o	2	4.8
t	9	10.8	a	1	3.4
g	6	7.6	p	1	2.3
c	5	8.9	n	0	1.8
i	5	10.0	e	0	1.6
u	5	7.8	m	0	1.2

CONCLUSIONS

The aim of this study was to collect data on user acceptance of pen interfaces experienced in simulated task contexts. Within these contexts we monitored both recognition accuracy and subjective evaluation. The general picture that emerges is that recogniser performance is a significant factor in determining user satisfaction, but that its impact depends on the nature of the task being performed, and the functional advantage of translating 'ink' traces into recognised text. This is essentially a cost/benefit relationship; users will accept the costs associated with recognition errors if (as in our record updating task) there is a substantial payoff in terms of achieving task goals. In this context, variations in recognition error rates that are generally within the range of 5-20% are only of minor importance in determining levels of user satisfaction. In tasks where there is a smaller benefit, users become more sensitive to the costs associated with handwriting recognition. In setting targets for recogniser performance, it is therefore inappropriate to think in terms of a fixed target that would make pen interfaces a viable option. It is rather the case that progressive increases in accuracy will extend the range of applications that users will find acceptable.

In the pursuit of improved recognition performance, we have also considered the relative contributions that might be made by users and system designers. One striking finding in this study was that recognition accuracy did not improve with practice. Given the immediate linkage between user behaviour and potentially rewarding system feedback, we might have expected to find some evidence of

behaviour modification. However, informal observation suggests that many individuals are either unwilling or unable to modify writing styles that involve stable, long-established motor skills. We might perhaps expect such changes to occur in the long term, but this assumes that users will find the pen system sufficiently attractive to reach this level of practice.

If users cannot adapt to the recogniser, the best (and widely used) strategy is to adapt the recogniser to the user. Analysis of the pattern of recognition errors for individual subjects indicates that in cost/benefit terms the best strategy will be to target the training procedure on a small subset of characters that are poorly recognised. This leaves open the question of how we can best identify this subset for individual users.

Finally, we should perhaps conclude by noting that most of our experimental subjects found the idea of a pen interface attractive. This enthusiasm was only slightly dimmed by their experience of the realities of unreliable recognition. Provided that pen applications are appropriately matched to recogniser performance, the future prospects for pen computing appear promising.

ACKNOWLEDGEMENT

REFERENCES

1. Santos, P.J., Baltzer, A.J., Badre, A.N., Henneman, R.L., and Miller, M.S. On handwriting recognition performance: Some experimental results. In: *Proceedings of the Human Factors Society 36th Annual Meeting*. Santa Monica CA: Human Factors Society (1992), pp 283-287.

2. Schoonard, J.W., Gould, J.D., Bieber, M., and Fusca, A. A behavioral study of a hand print recognition system. *IBM Research Report RC 12494*. Yorktown Heights, NY: T.J. Watson Research Center, 1987.

3. Wolf, C.G. Understanding handwriting recognition from the user's perspective. In: *Proceedings of the Human Factors Society 34th Annual Meeting*. Santa Monica CA: Human Factors Society, (1990), pp 249-253.

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

5. Tappert, C.C., and Jeanty, H.H. A study of several accuracy-improvement methods for a handwriting recognition system. *IBM Research Report RC 15373*. Yorktown Heights, NY: T.J. Watson Research Center, 1990.