

Human Performance on the USPS Database

Ibrahim Chaaban

Michael R. Scheessele

Abstract

We found that the human error rate in recognition of individual handwritten digits is 2.37%. This differs somewhat from two prior studies [1], [2].

INTRODUCTION

The recognition of handwritten digits is very challenging and it has been the subject of much attention in the field of handwriting recognition. Recognizing digits is a problem that at first seems simple, but it is non-trivial to program a computer to do it. The complexity of this task lies in the fact that a computer program must be able to recognize handwritten digits produced by different people, using different instruments. The system has to deal with widely different sizes and slants, with different shapes and widths of the strokes. Even so, with respect to individual handwritten digits, machine recognition systems have achieved an accuracy of 99.58% [3]. This invites the question, how does human performance compare to machine performance in this task? In fact, two experiments have been conducted to evaluate human performance in recognition of individual handwritten digits. Both experiments used the United States Postal Service (USPS) database. The USPS database contains 9298 handwritten digits divided into two subsets - a 'training' subset with 7291 digits and a 'test' subset with 2007 digits. Both human experiments used the 'test' subset. The first experiment reported a 2.5% error rate [1]¹, while the second experiment reported an error rate of 1.51% [2]. The first experiment was published as a proprietary report and is not readily available for public consumption. The second experiment suffered from two major methodological flaws. The experiment was conducted using four subjects and each was given 2007 test patterns

¹ We tried our best to locate a copy of this proprietary report, but we were unsuccessful. We contacted the authors, but neither were able to produce a copy of the report.

which were printed on white paper. Each page had approximately 120 images of handwritten digits separated by white space. Each subject was asked to identify each pattern and then to clearly label the pattern on the paper. The results were “carefully” entered into an Excel file manually [2]. The first flaw of the experiment is what we call association. When a subject looks at a sheet of paper that has 120 images on it, it is possible that when the subject encounters a difficult image to identify, this subject might associate this particular image with other images on the page in deciding on a response. The second flaw comes from the fact that the subjects wrote their responses on the papers they were given. How do we know that the person entering the results into Excel is reading the results (which are handwritten digits) correctly? It seems that the problem is being regenerated by the subjects, and now the person entering the results into Excel must solve the problem. Due to these two flaws in the design of the second experiment and the lack of availability of the report for the first experiment, we ran an experiment to determine human performance in recognition of individual handwritten digits.

EXPERIMENT

Method

Subjects

Four undergraduate IUSB students participated in this experiment². Subjects were at least 18 years old and had normal (20/20) or corrected-to-normal vision in both eyes. For completing the experiment each subject was paid \$30.00.

Stimuli

As in the prior studies, the ‘test’ subset (2007 digits) of the USPS database served as the stimuli. The USPS database was originally collected by CEDAR. Then it was modified

² This experiment was first approved by the IUSB Institutional Review Board.

by LeCun's research group [4]. The binary patterns were transformed into a 16×16 pixel box that kept the same aspect ratio and centered the patterns. The resulting patterns were gray-level and scaled and translated to fall within the range from -1 to 1. On our system this caused the digits to appear white against a black background. So we reversed the colors to match what would normally appear on an envelope (black text on white paper). The files containing the vectors of these images can be found at [4].

Procedure

Each subject attended 3 sessions of approximately one hour each. In the first two sessions there were 700 trials, and in the third session there were 607 trials, for a total of 2007 trials. In each session the subject sat in front of a computer screen and used software especially created for this experiment (Figure 1).

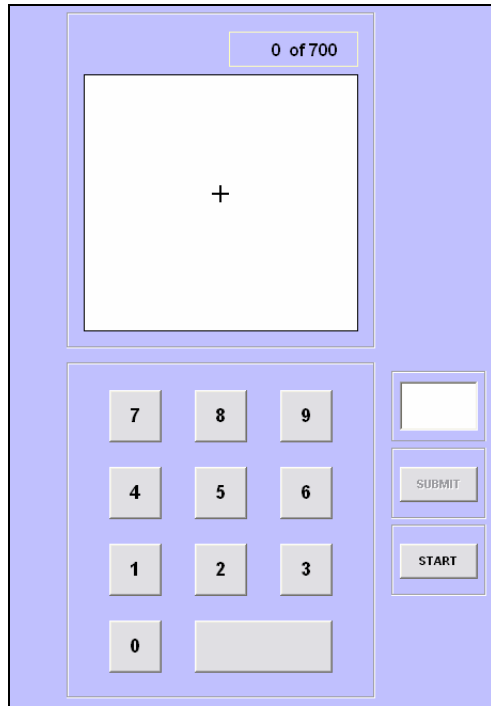


Figure 1. A screenshot of the software used for the Experiment.

A subject's task was to identify a series of handwritten digits randomly presented on the computer screen, using the mouse to respond. Subjects were asked to respond even if a stimulus was ambiguous. At the end of each session the software calculated the percent correct.

Results

As reported in the experiment of Dong et al. [2] there are four labeling errors in the USPS database (Table 1).





PATTERN NUMBER	16 × 16 IMAGE	USPS LABEL
234		1
971		4
994		5
1978		5

Table 1. This table shows the USPS misclassified patterns. The first column (leftmost) shows the pattern numbers, the middle column shows the actual 16×16 patterns, and the third column (rightmost) shows the USPS labels of the patterns. In each case, note that the USPS labeling appears to be incorrect.

Without taking those labeling errors into consideration, the average percent error was 2.57% (Table 2).

Subject #	Session 1 Correct Responses	Session 2 Correct Responses	Session 3 Correct Responses	Total Correct Responses	Percent Error
1	688	682	590	1960	2.34%
2	685	683	592	1960	2.34%
3	685	681	594	1960	2.34%
4	681	676	585	1942	3.24%
Total Number of Trials	700	700	607	2007	
Average Percent Error					2.57%

Table 2. Results obtained from the Experiment. The average percent error was 2.57%.

After removing the four incorrectly labeled digits, the average percent error rate was 2.37% (Table 3).

Subject #	Session 1 Correct Responses	Session 2 Correct Responses	Session 3 Correct Responses	Total Correct Responses	Percent Error
1	688	682	590	1960	2.15%
2	685	683	592	1960	2.15%
3	685	681	594	1960	2.15%
4	681	676	585	1942	3.05%
Total Number of Trials	699	698	606	2003	
Average Percent Error					2.37%

Table 3. Results obtained from Experiment. After removing the four trials with mislabeled images, the average percent error was 2.37%.

This is comparable to the error rate of 2.5% found by Bromley and Sackinger [1] and higher than the error rate of 1.51% reported by Dong et al. [2].

CONCLUSION

We found that human recognition of individual handwritten digits, using the USPS database 'test' subset is 97.63% (error rate: 2.37%). The slightly higher error rate reported by Bromley and Sackinger [1] apparently was due to their inclusion of the four incorrectly labeled digits in the USPS database. The lower error rate reported by Dong et al. [2] was apparently due to a flaw in their experimental design.

REFERENCES

- [1] Bromley, J., & Sackinger, E. (1991). Neural-network and k-nearest-neighbor classifiers. Tech. Rep. 11359-910819-16TM, AT&T.
- [2] Dong, J., Xiong, K. A., & Suen, C. Y. (2002). Statistical Results of Human Performance on USPS Database. Retrieved April 8, 2005, from <http://www.cenparmi.concordia.ca/people/jdong>
- [3] Liu, L. C., Nakashima, K., Sako H., & Fujisawa, H. (2002). Integrated Segmentation and Recognition of Handwritten Numerals: Comparison of Classification Algorithms. International Workshop on Frontiers in Handwritten Recognition (IWFHR), 8, 303-308.
- [4] United States Postal Service Database. Retrieved September 8, 2005, from <http://www.kernel.org/data.html>