

## RESULTS FROM A PERFORMANCE EVALUATION OF HANDWRITTEN ADDRESS RECOGNITION SYSTEMS FOR THE UNITED STATES POSTAL SERVICE

DONALD P. D'AMATO  
Mitretek Systems, Inc., 7525 Colshire Drive, McLean, VA, 22102-7400 USA  
E-mail: [ddamato@mitretek.org](mailto:ddamato@mitretek.org)

EDWARD J. KUEBERT AND ALFRED LAWSON  
United States Postal Service, 8403 Lee Highway, Merrifield, VA, 22082-8101 USA  
E-mail: [ekuebert@email.usps.gov](mailto:ekuebert@email.usps.gov), [alawson@email.usps.gov](mailto:alawson@email.usps.gov)

For a cost-incentive-based procurement (known as HIP), the U.S. Postal Service (USPS) developed a methodology to predict the recognition performance of Remote Computer Reader (RCR) systems for handwritten letter mail. Very high volumes of mail in the United States mean that slight changes in mail piece finalization and error rates have substantial cost consequences. Thus, high measurement precision and carefully truthed data are required. Because of considerable regional and seasonal variability in address quality, the HIP evaluation required large, representative databases of images and confirmation using high volumes of live-mail. At least four RCR versions were evaluated in HIP. In comparison to a baseline RCR system, the final HIP RCR system achieved the considerable gain of approximately 33 percent in the finalization rate for an image database, while reducing the error rate to about 2.5 percent. Live-mail measurements from 25 diverse sites corroborated the database results and illustrated the high variability in address quality and consequent recognition performance. USPS' testing confirmed that evaluation with sufficiently large and representative databases is an effective means for predicting performance on live-mail.

### 1 Background

The United States Postal Service (USPS) processes some 13.4 billion pieces of handwritten letter mail annually. It employs a combination of manual key-entry from digital images and multiline optical character recognition (OCR) to develop an 11-digit delivery point bar code<sup>1</sup> for encoding U.S. letter mail. The bar code, known as the POSTNET Code, is sprayed on the lower right front of an envelope, facilitating fully mechanized, highly reliable sorting of the mail to a carrier's walk sequence. The OCR algorithms, which encode most of the handwritten letter mail (and a smaller portion of the machine printed letter mail) reside in Remote Computer Reader (RCR) systems now deployed at 254 sites nationwide. Substantial computing power is required in the RCR

---

<sup>1</sup> Delivery point bar code: An 11-digit bar code representing the ultimate delivery point for which a mail piece is destined; comprising the ZIP+4 code (i.e., 9 digits) and, for regular residential mail, the last two digits of the household street address number.

systems to achieve throughputs that range from 30,000 to 330,000 mail pieces/hour, depending upon site size (i.e., letter mail volume and image lift capacity).

Recent advances in postal OCR systems in the United States have dramatically improved the percentages of handwritten letter mail that can be read fully automatically. Prior to 1997, the percentage of handwritten letter mail readable by USPS OCR systems was only about two percent. Thirteen years of USPS-supported research conducted at the State University of New York (SUNY) at Buffalo<sup>2</sup> resulted in commercially-produced equipment that increased this percentage to about 23 percent.

In 1997, the USPS initiated its Handwritten Improvement Program (HIP) to raise substantially the finalized encode rate (finalization rate) for handwritten letter mail on its RCR systems. In HIP, a USPS contractor integrated within a highly parallel, multiprocessor system OCR algorithms recently developed at SUNY-Buffalo and by a for-profit OCR firm to achieve the recognition speed and accuracy goals.

A mail piece is “finalized”<sup>3</sup> when no further encoding is possible, as determined by the RCR system itself. The total finalization rate is equal to the sum of percentages of input images finalized to three levels of sortation:

- 5-digits (i.e., to the Post Office level)
- 9-digits (i.e., to a side of a city block, a firm, or a building)
- 11-digits (i.e., to the carrier’s delivery point sequence)

The initial goal of HIP was to reach within two years a 50 percent total finalization rate for handwritten letter mail, while simultaneously maintaining or reducing the error rate, which was at that time slightly less than three percent of the finalized pieces. Engineers and contracting officers at the USPS devised an incentive contracting system that envisioned payments to the HIP contractor based on measured percentage point improvements in the total handwritten letter mail finalization rates, rather than on costs incurred by the contractor. The HIP contractor bore the expense of deploying all hardware and providing the complete support infrastructure to ensure a predetermined throughput at each site. The USPS would not incur payment liability if there was no finalization rate improvement, even if hardware was deployed to maintain a site’s throughput while attempting to raise the encode rate.

---

<sup>2</sup> S. N. Srihari and E. J. Kuebert, Integration of Hand-Written Address Interpretation Technology into the United States Postal Service Remote Computer Reader System, Proc. 4<sup>th</sup> ICDAR, Ulm, Germany, pp 892-896, 1997.

<sup>3</sup> The number of input images “finalized” is not the same as the number of mail piece images correctly recognized. A mail piece can be finalized, yet be incorrectly encoded.

The USPS has verified that the RCR systems procured through HIP achieved a 57 percent average total finalization rate within the two-year period, while reducing the error rate to 2.49 percent. As a result, the USPS saved many millions of work-hours of key-entry labor and thus far, eleven of its Remote Encoding Centers for key-entry have been closed or are scheduled for closure.

## 2 Methodology

USPS management developed a methodology to ascertain that the USPS received the contracted rate improvements. The need for self-oversight was magnified by the fact that the very high volumes processed mean that slight changes in the finalization and error rates have significant consequences for the USPS' operating costs. The USPS worked with Mitretek Systems, Inc. (a not-for-profit, conflict-free, systems engineering corporation) concerning its computations of the encoding rates and verification of statistical significance of the measured gains.

There is considerable variability in the quantity and quality of the handwritten address images processed by the RCR systems, both among the sites and at a single site over time. Some factors that can affect the address quality are the following:

- Daily, weekly, monthly, and seasonal variability in the mail stream
- Varying modes of operation of systems that capture the input images (the AFCS<sup>4</sup> and MLOCR<sup>5</sup>), depending upon the availability of labor and/or equipment
- Varying frequency and quality of AFCS and MLOCR preventative maintenance procedures
- Differing addressing schemes in each locality, including grid addressing, high-rise addressing, hyphenated-numeric addressing, and addresses with foreign words

An accurate analysis of an RCR system's potential cost savings requires that the test data reflect this considerable variability and that all mail characteristics be fairly represented through a suitably randomized sampling process. Therefore, USPS engineers and economic analysts developed a testing methodology to ensure that, within reasonable tolerances, the RCR systems met or exceeded expectations following deployment. A three-step testing process was devised, consisting of the following:

---

<sup>4</sup> AFCS: Advanced Facer Cancellor System. A machine that faces, cancels, and sorts incoming letter-size mail to one of seven stackers. An input subsystem modification provides the AFCS with image lifting capability.

<sup>5</sup> MLOCR: Multiline Optical Character Recognition system. An optical character reader that reads and interprets more than one line of the delivery address.

- Step I: A test would be run using one million contemporaneously collected images from handwritten letter mail. Five hundred thousand of these images (known as Sample Set 1) would come from the ten sites with the highest processing volumes and from the ten sites considered to have the most difficult addressing situations. Approximately 25,000 images were collected randomly from each of these 20 sites. The remaining 500,000 images (known as Sample Set 2) would come from a random selection of 100 sites, with approximately 5,000 images collected per site. To measure the error rate, 100,000 of these images (50,000 from each set) were ground-truthed using key-entry techniques.
- Step II: Twenty-five sites were selected as verification sites. The 50,000 contemporary images from each of the 25 sites were combined to create a test set of 1.25 million images (known as Sample Set 3). This set would be used to predict the finalization rate anticipated for each site during Step III. Ten of these sites were represented in Sample Set 1 and eight were represented in Sample Set 2, although the Sample Set 3 data were distinct from those in Sample Sets 1 and 2.
- Step III: A 30-day pre- and post-deployment test would be run with live-mail at each of the Step II sites to ensure that the USPS actually achieved the finalization rate improvements anticipated. The criterion employed was that the results of this set (Step III) should agree with those from Step II with less than a 10 percent probability of error.

### *2.1 The Postal Cost Model*

In automated postal address recognition systems, many factors must be considered to determine an overall confidence in the determination of the delivery point sequence bar code. The development of an address recognition system involves selecting many thresholds to determine an optimum tradeoff point between the percentage of addresses correctly recognized and those in error. The tradeoff point should be selected by the system developer to result in the lowest overall mail-processing cost.

To predict the cost consequences for each specific set of performance measurements, the USPS developed a Cost Model. The processing costs associated with differing levels of sortation and categories of recognition errors were calculated using detailed postal flow models that trace the paths of letters from initial image capture to their final destinations. The recognition results using the truthed images of Step I served as the source of input data to the Cost Model. The Cost Model is actually a matrix of the cost consequences for specific combinations of address potentials (the

finest levels of sort to which mail pieces could be resolved using the available information) and outcomes from the RCR processing.

## *2.2 The Baseline System*

For its performance comparisons, the USPS selected a previous version of its RCR system to use as a Baseline. Percentage improvements above the Baseline's finalization rate determined the incentive payments to the HIP contractor. That version, known as Version 4.03, was achieving the aforementioned 23 percent total finalization rate for handwritten letter mail when HIP was initiated.

It should be noted that the denominator (i.e., the number of input images that are handwritten) in the RCR system's calculation of its handwritten address finalization rates is, itself, system-determined. For this reason, a comparative evaluation of handwritten address recognition systems requires that a "Baseline" determination be made of the number of addresses in the input database which are handwritten and that, whenever possible, this number should be used in lieu of the system-determined quantity. Otherwise, the system's determination of its handwritten finalization rate could be improved by its categorizing some of the more difficult input images as being non-handwritten (i.e., as machine-printed).

## *2.3 The HIP Versions*

Under HIP, three versions of the RCR software were delivered and evaluated, namely:

- Version 5.0, first tested by the USPS in-house at its Engineering Facility in Merrifield, Virginia during July 1998
- Version 5.1, first tested in-house during October 1998
- Version 7.1.1, first tested in-house during July 1999 and deployed for the processing of live-mail at initial field sites during August 1999

## *2.4 Statistical Analysis Program*

To provide a consistent and straightforward means to input RCR performance data and to compare finalization rates achieved by several releases of the RCR system software, Mitretek developed a database program—its Statistical Analysis Program—using Microsoft Access 97 and Visual Basic for Applications. The Statistical Analysis Program uses two different sources of input data produced by the RCR software—summary sheet text files and continuously updated binary files, known as RCRSTATS files—and provides various types of statistical calculations for the image databases and live-mail. For more sophisticated statistical analyses, a commercially available software package that works in conjunction with Microsoft Access was purchased.

The results from testing the four RCR versions with the image databases at the Merrifield, Virginia facility were incorporated into the Statistical Analysis Program database. For live-mail, the results from over 14 system-years of pre-and post-deployment testing, obtained from more than 125 5-Megabyte RCRSTATS files, were aggregated hourly, daily, and weekly, and then imported into the database.

### 3 Steps I and II Results

Figure 1 displays the handwritten address finalization rates for the four RCR versions—Versions 4.03 (the Baseline), 5.0, 5.1, and 7.1.1—for the three separate Sample Sets and Combined Sample Sets 1 and 2. The bar heights represent weighted mean finalization rates and the surrounding error bars represent standard errors of the means. Note that the vertical axis is scaled from 10 to 70 percent. The error bars are tightest for the Combined Sample Sets because the highest number of sites is represented in the Combined Sets.

As may be observed, the results for Sample Sets 2 and 3 are within the standard errors of one another, while the results for Sample Set 1 are somewhat lower. It is suspected that this occurred because Sample Set 1 was obtained from sites having higher volumes and more difficult addressing schemes; whereas, Sample Sets 2 and 3 are probably more representative of all sites.

The total finalization rate results measured for Combined Sample Sets 1 and 2 are:

- 22.36%  $\pm$  0.40% for Version 4.03
- 55.01 %  $\pm$  0.59% for Version 7.1.1

The result is a very substantial gain of 32.65%  $\pm$  0.37% for the final Version of HIP over that achieved by the Baseline Version.

For Combined Sample Sets 1 and 2, figure 2 displays the histograms of the finalization rates measured for the four RCR versions, with the values grouped into intervals of two percentage points. The distributions for the four RCR versions appear to be similarly shaped and of roughly equal width, with only a few outliers on their lower sides. The site with the lowest total finalization rate in each of the distributions is the same site and is located within a major urban area.

As a part of the Step I testing, the USPS measured the error rates for the 100,000 truthed images. The HIP contract required that the error rate be maintained at less than three percent of the finalized mail pieces. The HIP contractor was able to achieve this result at the finalization rates reported.

#### 4 Step III Results

Step III consisted of measuring live-mail pre- and post-deployment finalization rates at the first 25 deployment sites, each over a period of at least 30 days.

The pre-deployment (Version 4.03) and post-deployment (Versions 5.0, 5.1, 7.0,<sup>6</sup> and 7.1.1) RCRSTATS binary data files were used to compute the total finalization rate improvements for each Step III site. Each RCRSTATS file contains at least 42 days of performance information, with data collected in five-minute intervals. Unlike the analyses for the image databases, the denominators in the calculations of the live-mail finalization rates had to be machine-determined, with no feasible procedure to determine independently the number letter mail pieces that were handwritten.

For each of the 25 sites, figure 3 displays the measured Step III finalization rates for RCR Versions 4.03, 5.0, 5.1, 7.0, and 7.1.1. The set of points on the right side of the graph represent the averages (unweighted by mail volumes) for these five versions. To the right of the graph, the corresponding numerical values are provided. For these 25 sites, the average finalization rate is  $23.25\% \pm 0.55\%$  for the Baseline Version and  $59.95\% \pm 0.55\%$  for Version 7.1.1, thus producing a gain of  $36.70\% \pm 0.85\%$ . Even if the differing denominators in the respective calculations is accounted for, this live-mail result represents a somewhat higher finalization value than that achieved using the image databases. Some of this difference may be attributed to the fact that the national address directory for the image database testing was selected at the beginning of the HIP contract and remained unchanged during the evaluation. In contrast, address directories used during the live-mail testing were typically updated weekly.

For Version 7.1.1, it is observed that (with the exception of Processing Center I, which represents the results from less than a single day) Processing Center E has the highest total finalization rate (at an average of 66.37 percent during 55 days), while Processing Center N has the lowest rate (at an average of 51.42 percent during 40 days.) The predominance of high-rises and numeric street addressing within the Processing Center N area probably account for its diminished rate.

Figure 4 displays this live-mail data as five histograms, again with the values grouped into intervals of two percentage points. As was the case for the Sample Sets, the distributions for live-mail appear to be similarly shaped and of roughly equal width, with just a few outliers on their lower sides. The distributions are more narrow than

---

<sup>6</sup> According to the HIP Contractor, Version 7.0 used the same recognition algorithms and parameters as Version 5.1. The measured performance differences for live-mail between Versions 5.1 and 7.0 probably reflect seasonally differing mail characteristics.

those for the Sample Sets because the number of images processed per site is substantially larger and the sites are fewer in number and less diverse.

## **5 Variability in the Mail Stream**

As previously mentioned, there is considerable variability in the quality of handwritten mail's addresses and images. Figure 5 illustrates some of this variability. In the figure, handwritten letter mail volume and total finalization rate are plotted hourly over the course of one week for the Processing and Distribution Center (P & DC) with the highest volume of mail. At this site, there is a fairly consistent daily pattern in volume and finalization rate, with hourly volume slowly increasing after about 7:00 AM, peaking at about 5:00 PM, and remaining steady until about 11:00 PM. The exceptional days are Sunday and Monday, with Sunday's volume substantially lower than average and Monday's volume higher, but more variable. Total finalization rate rises more quickly each day and remains steady, at just under 60 percent, until after midnight.

To understand better hour-to-hour variability in the mail stream, total handwritten address finalization rates obtained from Version 7.1.1 RCRSTATS files from the 25 Step III sites were aggregated hourly and plotted as a histogram. Figure 6 displays this histogram, using one percentage point increments, and represents the processing of almost 140 million handwritten addresses. The distribution's unweighted mean is 53.6 percent; the mean, weighted by the handwritten address volume processed in each hour, is 59.3 percent; and the distribution's mode is about 62 percent.

The distribution is remarkably wide, with an unweighted sample standard deviation of about 16 percent —illustrated by the horizontal dotted line. Even when weighted by volumes processed, the sample standard deviation—illustrated by the horizontal solid line—is almost 10 percent. Instances with a zero finalization rate occurred during slack processing periods and actually represent a very low number of mail pieces.

## **6 Conclusions**

The USPS believes that HIP achieved its objectives and the reported results demonstrate that testing with sufficiently high-volume image databases is an effective means for predicting the recognition performance with live-mail. The finalization rates achieved during live-mail testing confirmed the predictions from the database testing.

The use of a Baseline version for performance comparisons to determine the incentive payments seems fair to system developers and to the USPS, although new baselines and databases will be required as systems and addressing schemes evolve.

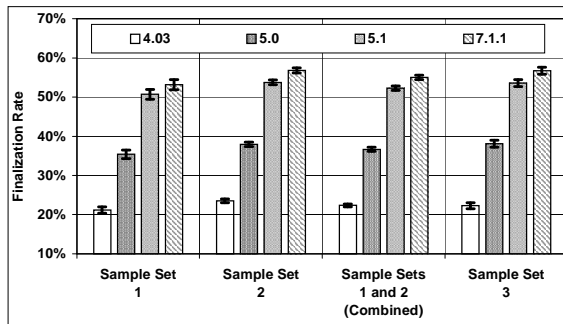


Figure 1. Comparison of total handwritten finalization rates for Sample Sets 1, 2, and 3

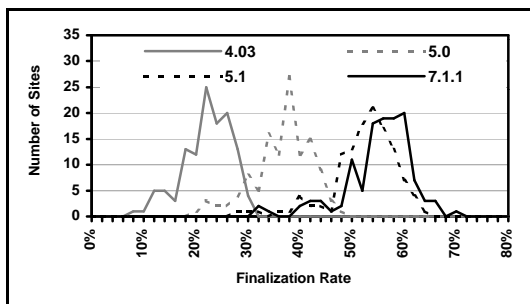


Figure 2. Histograms of total handwritten finalization rates for Sample Sets 1 and 2 (Combined)

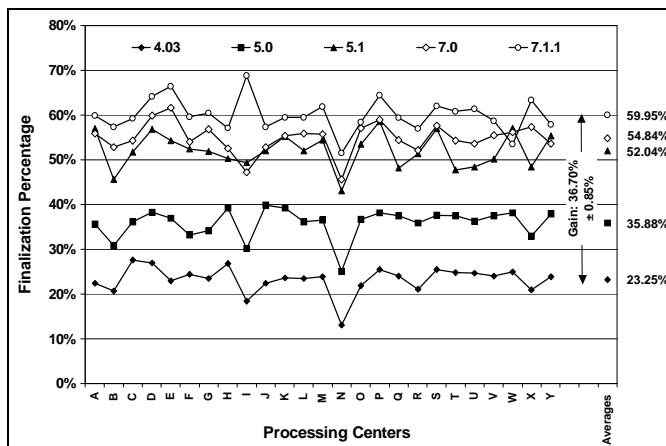


Figure 3. Live-mail finalization rates for 25 Step III P & DCs

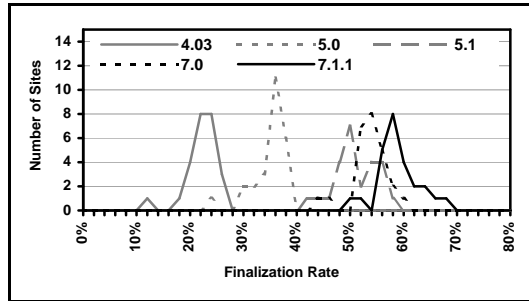


Figure 4. Histograms of total handwritten finalization rates for live-mail

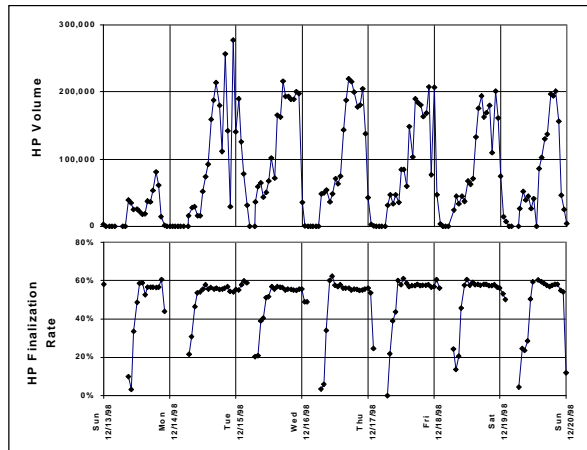


Figure 5. Hour-to-hour handwritten letter mail volume and finalization rate during one week in December 1998 at a single P & DC (using RCR Version 5.1)

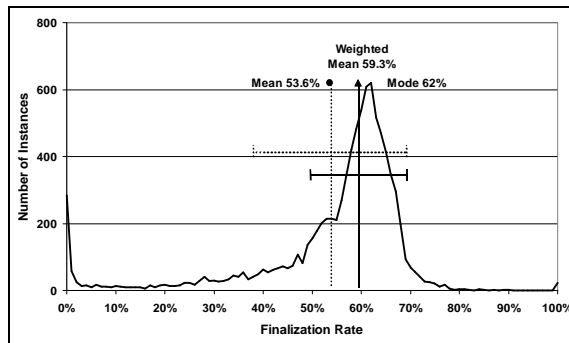


Figure 6. Histogram of the total handwritten finalization rates for RCR Version 7.1.1 recorded hourly at 25 P & DCs