AD-A242 516

DTIC
ELECTE
OCT 31 1991
S C D

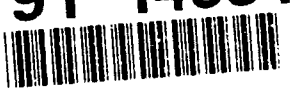
210600-43-T

# ADVANCED RESEARCH IN CONTEXTUAL ANALYSIS OF ADDRESSES: PHASE III REPORT

Andrew M. Gillies
Alan J. Vayda
Daniel J. Hepp
Michael A. Janeczko

June 1991
(Draft version: April 1991)

91-14581



Submitted to:
U.S. Postal Service
Office of Advanced Technology
Technology Resource Department
Washington, D.C. 20260-8121

Contract Number: 104230-86-H-0042
Task Number: 104230-88-D-2575

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

ERIM
P.O. Box 134001
Ann Arbor, MI 48113-4001

U. S. Postal Service

# REPORT ABSTRACT

Report Number [redacted]

1. Submitting U.S. Postal Service Organization

Technology Resource Department Office of Advance Technology

2. Library Liaison Officer

Susan Schaeffer

3. Telephone Number

202-268-3867

4. Date Form Completed

April 17,1991

5. Report Title

Advanced Research in Contextural Analysis of Addresses: Phase III Report

6. Report Date

April 1991

7. Organizational Author (Requiring Group or Department, Ad Hoc Task Force, or Contractor, as applicable)

Environmental Research Institute of Michigan

8. Contract Number

Contract: 104230-86-H-0042 Task: 104230-88-D-2575

9. Restrictions

No Restrictions X Official Use Only — Limited Official Use Only —

10. Contact (Personal Author, Project Manager, Project Director, etc.)

Andrew M. Gillies

11. Telephone Number

313-994-1200

12. Descriptors
(Subjects, Keywords, etc.)

| | |
|---|---|
| Contextual Analysis | Segmentation |
| Address Block Interpretation | Optical Character Recognition |
| Word Recognition | |
| Word Verification | |
| Directory Matching | |

13. Abstract of Report

This report describes the continued development and testing of a system for contextual analysis of machine printed address block images. The system receives a binary image of the address block (location of the address block is not a part of this work) and then: 1) segments the image into lines, words, and characters with multiple hypotheses, 2) assigns class confidence to each character hypothesis using neural networks, 3) locates, reads, and reconciles the city name and ZIP code, 4) parses the address block using keyword recognition, 5) if a PO Box is found, reads the box number and verifies it against the postal directory, otherwise, 6) forms a street name lexicon based on contextual information, including number of street name words, word lengths, recognition of suffix and directionals, and the ZIP code, 7) forms an additional street name lexicon based on partial recognition of the street words, 8) uses word recognition within these lexicons to rank street name hypotheses, 9)retrieves street and range records from a postal directory, 10) matches information from the retrieved records to the fields on the mailpiece forming 9-digit ZIP code hypotheses, 11) applies decision logic to assign the finest supportable depth of sort. In an end-to-end test on data selected for OCR difficulty, using corrected LOS scoring, the system had an encode rate of 50% (with 9.5% error) and an accept rate of 84% (with 9.3% error). This compares favorably with an encode rate of 16.7% (with 13.6% error) and an accept rate of 61% (with 15.5% error) achieved by the current MLOCR machine on this same dataset. In related tests, the word recognition submodule performed at 96% with a lexicon size of 100, and 92% with a lexicon size of 2000. A detailed analysis of errors and rejects is included in the report. This analysis includes a study of patron errors. A plan for further work is also included.

## TABLE OF CONTENTS

Accession For
NTIS GRA&I
DTIC TAB
Unannounced
Justification
By Per Form 50
Distribution/
Availability Codes
Dist | Avail and/or Special
A-1

## List of Figures

## List of Tables

# 1. INTRODUCTION

This report describes ERIM's contextual analysis system and presents the results of tests performed at the end of Phase III of the project.

The balance of this report briefly describes the workings of the contextual analysis system and gives details and analysis of the test results. Section 2 gives a brief system description. Section 3 details the results of the word verification test. Section 4 details the end to end system test results. Section 5 gives a detailed analysis of errors in the end to end system. Section 6 contains an examination of the computation time requirements of the system. Section 7 contains an analysis of the patron errors detected in the test data. Section 8 contains a plan for Phase IVa of the project. Finally, Section 9 provides a brief discussion and conclusions.

# 2. SYSTEM DESCRIPTION

The major modules of the Phase III contextual analysis system are depicted in Figure 2.1. This diagram represents the data-flow through the system. Each module uses the results of previous modules (as shown by arrows) and also, as results accumulate, the results from modules further back in the chain. The word verification and database systems, not shown in the diagram, are used by several modules in the system. The working of the each module in the end to end system will be described briefly in the following sections.

## 2.1. Image Segmentation

The image segmentation module performs the following steps. Image tilt correction, where initial tilt is determined by Hough transform. Segmentation of the address block into lines. Segmentation of each line into words with multiple hypotheses. Segmentation of each word into characters with multiple hypotheses.

## 2.2. Fuzzy Logic City/ZIP Reader

The fuzzy logic City/ZIP reader assigns a 5 digit and/or 3 digit (city default) ZIP code to the mailpiece. This module uses its own parsing with loose constraints on word arrangement to locate city, state and ZIP code word images. The system looks at each city name with a high word verification value. It then uses word verification for all legal ZIP codes in that city to assign a ZIP code confidence. It then assigns the ZIP code using fuzzy logic. The fuzzy logic rule returns the ZIP code which maximizes the score of the minimum between the city name confidence and the ZIP code confidence for each legal pair. A similar system is also used to assign city default ZIP codes.

## 2.3. Address Block Parser

The address block parser uses models of address block formats to assign roles to words in the address block image. Multiple parses are generated and ranked according to confidence. Confidence is assigned using number versus non-number recognition, word verification of suffix, directional, state, and other key words.

## 2.4. Numeral Reading System

The numeral reading system uses neural networks to read the characters in the ZIP code word assigned by the top ranked parse from the address block parsing module. The module outputs a 3 and a 5 digit version for each ZIP code, and a 9 digit version if 9 digits are present.

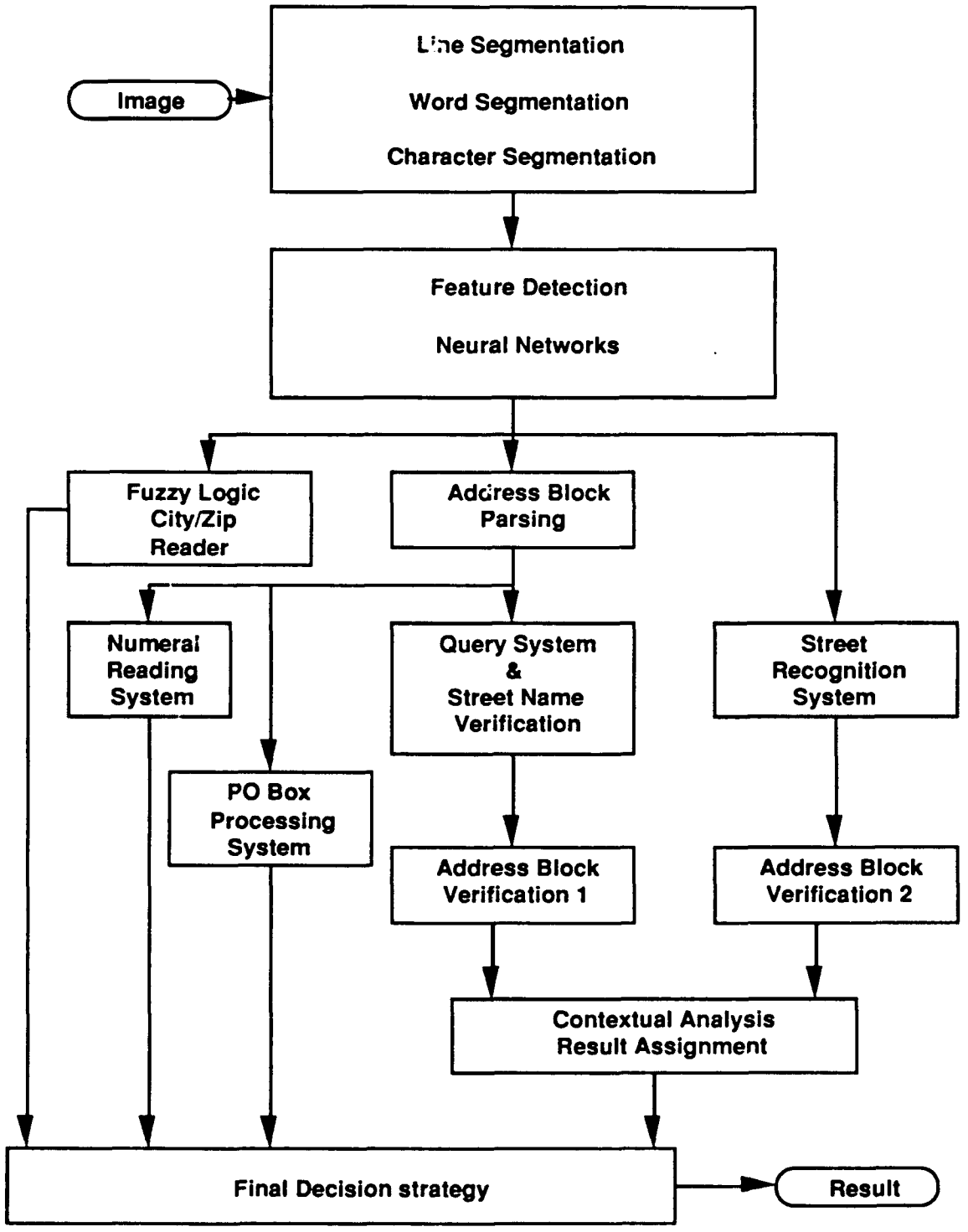



Figure 2.1 - Overview of system modules.

## 2.5. Query System

The query system uses the results from address block parsing to generate queries into the database. The street names from records matching the queries are verified, and ranked according to confidence. This module generates queries from the 7 top-ranked parses. Each parse can give rise to several alternate queries due to multiple hypotheses from word-length and key word verification systems. The module also uses backoff rules which form more general queries by leaving out selected pieces of information from previously generated queries. The output of the query and verification system is a list of candidate street names.

## 2.6. Street Recognition System

The street recognition system (formerly referred to as the system 2 hypothesis generator) represents an alternate approach to generating street name hypotheses. The module uses its own parsing with loose word arrangement constraints to locate potential street name word images. It then uses neural network character recognition and a 2-out-of-4 matching criterion to select street name words from the entire 4-SCF word list. It then queries the database for street names containing those words and verifies and ranks street names.

## 2.7. PO Box Processing System

The PO Box processing system handles assignment of 9 digit codes from post office box records. It looks at the top ranked parse from the address block parsing system to see if it contains a post office box number. If so, it reads the number and tries to assign a 9 digit code based on the box number, city name, and ZIP code.

## 2.8. Address Block Verification 1

The address block verification 1 module tries to assign a 9 digit code based on the top ranked street name from the query and street name verification system. The module looks at database records containing the hypothesized street name. Based on the number of such records it may also narrow its search based on additional information from the query which produced the hypothesis. The module examines fields in the database records and attempts to match those fields to elements from the image. It generates a ranked list of matching database records.

## 2.9. Address Block Verification 2

The address block verification 2 module tries to assign a 9 digit code based on the result of the system 2 hypothesis generator. It investigates several top ranked street names from the hypothesis list generated by system 2. As with the address block verification 1 module, a ranked list of matching database records is generated.

## 2.10. Contextual Analysis Result Assignment

The contextual analysis result assignment system combines the results from the two address block verification systems to arrive at a lowest expected cost 9-digit (or 5 digit) ZIP code assignment. The system uses information about the hierarchy of database records to discover cases where 2 (or more) conflicting records match with approximately equal confidence. In these cases, the decision passes up the hierarchy to a point where the matches are consistent. The module also weighs the processing costs and against confidence values to arrive at the decision with the lowest expected cost.

## 2.11. Final Decision Strategy

The final decision strategy integrates the output from four modules: 1) the fuzzy logic city/ZIP reader, 2) the numeral reading system, 3) the PO box processing system, and 4) the contextual analysis result assignment system. It generates the final output from the end to end system. The final decision is arrived at based on an ordering of the outputs from the various modules. At each stage in the order the appropriate module is interrogated to see if generates a certain kind of answer with acceptable high confidence. If so that result is taken. If not, the processing passes to the next step. The order of processing is as follows:

1) Numeral reading system generates 5 digit direct.
2) Fuzzy logic city/ZIP reader generates 5 digit direct.
3) POB system generates a 9 digit result.
4) Contextual analysis result assignment to 9 or 5 digits.
5) Fuzzy logic city/ZIP reader generates 5 digit result.
6) Fuzzy logic city/ZIP reader generates 3 digit result.
7) Numeral reading system generates 9 digit result.
8) Numeral reading system generates 5 digit result.
9) Numeral reading system generates 3 digit result.
10) Reject the mailpiece.

## 3. WORD VERIFICATION TEST RESULTS

The word verification module assigns confidence to a given word for a given word image. For Phase III, the word verification module system uses multiple-hypothesis character segmentation and neural networks to assign confidence to a given word/word-image pair.

The word verification module was tested on 1055 word images. For this test 20 lexicons varying in size from 100 to 2000 strings were used. For each lexicon size, the results were averaged among the lexicons of that size. For the most difficult case, a 2000 string lexicon, 91.8% of the 1055 word images had their correct string appearing as the top ranked candidate. For lexicons of 100 strings each, the percent achieving top rank was 96.3%. The results of this test are summarized in Table 3.1.

Table 3.1 - Word Verification Results

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 91.8 | 93.8 | 94.6 | 95.0 | 95.2 | 95.6 | 95.6 | 95.9 | 95.9 | 96.1 | 100.0 |
| 1000 ( 2) | 93.4 | 94.8 | 95.3 | 95.7 | 96.0 | 96.3 | 96.6 | 96.9 | 97.2 | 97.3 | 100.0 |
| 500 ( 2) | 94.4 | 95.9 | 96.4 | 96.9 | 97.3 | 97.3 | 97.4 | 97.6 | 97.7 | 97.7 | 100.0 |
| 200 ( 5) | 95.5 | 96.9 | 97.2 | 97.5 | 97.6 | 97.8 | 97.9 | 98.0 | 98.1 | 98.2 | 100.0 |
| 100 (10) | 96.3 | 97.4 | 97.8 | 98.0 | 98.2 | 98.3 | 98.4 | 98.5 | 98.5 | 98.5 | 100.0 |

The character segmentation module presents the word verification module with multiple segmentations of the word image into characters. These word segmentations are generally of different lengths. We can use the average of these lengths as an estimation of word length. Table 3.2 shows the percentage of word images whose true length differs from this estimate by a specified amount.

Table 3.2 - Word Length Estimation Results

| | |
|---|---|
| Correct | 29.0 |
| Off by One (±1) | 52.9 |
| Off by Two (±2) | 16.5 |
| Off by Three (±3) | 1.3 |

Of the 1055 word images, 6, or 0.57%, resulted in segmentations where the correct length was not among the ones generated by the character segmentation module.

The following three tables show the word verification results broken down by image sortation category (5, 9, or R). The cardinality shown with each table gives the number of word images falling into the category.

Table 3.3 - SET 5 (Cardinality: 642)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 94.9 | 96.4 | 97.4 | 97.5 | 97.5 | 98.0 | 98.0 | 98.3 | 98.3 | 98.4 | 100.0 |
| 1000 ( 2) | 96.2 | 97.3 | 97.7 | 98.1 | 98.4 | 98.7 | 98.8 | 98.9 | 99.1 | 99.1 | 100.0 |
| 500 ( 2) | 96.9 | 98.3 | 98.5 | 98.8 | 99.1 | 99.1 | 99.1 | 99.2 | 99.2 | 99.2 | 100.0 |
| 200 ( 5) | 97.8 | 98.8 | 99.0 | 99.1 | 99.1 | 99.1 | 99.2 | 99.2 | 99.2 | 99.2 | 100.0 |
| 100 (10) | 98.4 | 99.0 | 99.2 | 99.2 | 99.2 | 99.3 | 99.3 | 99.3 | 99.3 | 99.3 | 100.0 |

Table 3.4 - SET 9 (Cardinality: 80)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 95.0 | 95.0 | 95.0 | 95.0 | 95.0 | 95.0 | 95.0 | 95.0 | 95.0 | 95.0 | 100.0 |
| 1000 ( 2) | 95.0 | 95.0 | 95.0 | 95.0 | 95.0 | 95.0 | 95.6 | 95.6 | 96.3 | 96.3 | 100.0 |
| 500 ( 2) | 95.0 | 95.6 | 95.6 | 95.6 | 96.3 | 96.3 | 96.3 | 96.3 | 96.3 | 96.3 | 100.0 |
| 200 ( 5) | 95.3 | 96.0 | 96.0 | 96.0 | 96.3 | 96.3 | 96.3 | 96.3 | 96.3 | 96.3 | 100.0 |
| 100 (10) | 95.6 | 96.1 | 96.3 | 96.3 | 96.3 | 96.4 | 96.9 | 96.9 | 97.0 | 97.1 | 100.0 |

Table 3.5 - SET R (Cardinality: 333)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 85.3 | 88.6 | 89.2 | 90.1 | 90.7 | 91.3 | 91.3 | 91.6 | 91.6 | 91.9 | 100.0 |
| 1000 ( 2) | 87.7 | 89.9 | 90.7 | 91.4 | 91.7 | 91.9 | 92.6 | 93.2 | 93.8 | 94.1 | 100.0 |
| 500 ( 2) | 89.5 | 91.3 | 92.3 | 93.7 | 94.0 | 94.1 | 94.4 | 94.9 | 95.0 | 95.0 | 100.0 |
| 200 ( 5) | 91.1 | 93.5 | 94.2 | 94.8 | 95.1 | 95.5 | 95.9 | 96.3 | 96.5 | 96.6 | 100.0 |
| 100 (10) | 92.3 | 94.5 | 95.5 | 96.1 | 96.5 | 96.7 | 96.9 | 97.1 | 97.2 | 97.3 | 100.0 |

The following tables show the word verification results broken down by word case.

Table 3.6 - UPPER-CASE (Cardinality: 631)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 92.7 | 94.6 | 95.4 | 95.9 | 96.0 | 96.4 | 96.4 | 96.7 | 96.7 | 96.8 | 100.0 |
| 1000 ( 2) | 94.3 | 95.7 | 96.3 | 96.5 | 96.7 | 96.8 | 97.3 | 97.5 | 98.0 | 98.2 | 100.0 |
| 500 ( 2) | 95.3 | 96.8 | 97.3 | 97.9 | 98.2 | 98.3 | 98.3 | 98.5 | 98.6 | 98.6 | 100.0 |
| 200 ( 5) | 96.4 | 97.8 | 98.2 | 98.4 | 98.5 | 98.6 | 98.7 | 98.9 | 98.9 | 99.0 | 100.0 |
| 100 (10) | 97.1 | 98.3 | 98.7 | 98.8 | 99.0 | 99.0 | 99.1 | 99.1 | 99.1 | 99.2 | 100.0 |

Table 3.7 - MIXED-CASE (Cardinality: 424)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 90.6 | 92.7 | 93.4 | 93.6 | 93.9 | 94.6 | 94.6 | 94.8 | 94.8 | 95.0 | 100.0 |
| 1000 ( 2) | 92.1 | 93.4 | 93.8 | 94.6 | 95.0 | 95.4 | 95.5 | 95.9 | 96.0 | 96.0 | 100.0 |
| 500 ( 2) | 93.0 | 94.6 | 94.9 | 95.4 | 95.9 | 95.9 | 96.1 | 96.3 | 96.3 | 96.3 | 100.0 |
| 200 ( 5) | 94.2 | 95.5 | 95.8 | 96.1 | 96.3 | 96.5 | 96.7 | 96.8 | 96.9 | 96.9 | 100.0 |
| 100 (10) | 95.0 | 96.0 | 96.4 | 96.7 | 96.9 | 97.1 | 97.2 | 97.5 | 97.5 | 97.6 | 100.0 |

The following tables show the word verification results broken down by word length.

Table 3.8 - LENGTH 2 (Cardinality: 122)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 77.9 | 83.6 | 87.7 | 89.3 | 90.2 | 91.0 | 91.0 | 91.0 | 91.0 | 91.8 | 100.0 |
| 1000 ( 2) | 83.2 | 88.5 | 89.8 | 90.6 | 91.4 | 91.8 | 93.0 | 94.3 | 95.9 | 96.7 | 100.0 |
| 500 ( 2) | 86.1 | 91.0 | 93.4 | 95.1 | 96.3 | 96.7 | 97.1 | 98.4 | 98.8 | 98.8 | 100.0 |
| 200 ( 5) | 89.8 | 95.4 | 97.0 | 98.4 | 98.5 | 98.9 | 99.3 | 99.7 | 99.8 | 100.0 | 100.0 |
| 100 (10) | 93.2 | 97.5 | 99.1 | 99.6 | 99.9 | 99.9 | 99.9 | 100.0 | 100.0 | 100.0 | 100.0 |

Table 3.9 - LENGTH 3 (Cardinality: 80)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 92.5 | 93.8 | 93.8 | 96.3 | 96.3 | 97.5 | 97.5 | 98.7 | 98.7 | 98.7 | 100.0 |
| 1000 ( 2) | 93.1 | 95.0 | 97.5 | 98.1 | 98.7 | 98.7 | 99.4 | 99.4 | 100.0 | 100.0 | 100.0 |
| 500 ( 2) | 95.0 | 96.9 | 96.9 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 200 ( 5) | 97.3 | 99.5 | 99.8 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 100 (10) | 98.3 | 99.8 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Table 3.10 - LENGTH 4 (Cardinality: 202)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 90.6 | 93.1 | 94.1 | 94.1 | 94.1 | 94.6 | 94.6 | 94.6 | 94.6 | 94.6 | 100.0 |
| 1000 ( 2) | 93.1 | 94.1 | 94.3 | 94.8 | 94.8 | 95.0 | 95.5 | 95.5 | 96.0 | 96.0 | 100.0 |
| 500 ( 2) | 93.8 | 95.3 | 95.5 | 95.5 | 96.0 | 96.0 | 96.0 | 96.0 | 96.0 | 96.0 | 100.0 |
| 200 ( 5) | 94.7 | 95.7 | 95.8 | 95.8 | 96.0 | 96.2 | 96.4 | 96.5 | 96.5 | 96.5 | 100.0 |
| 100 (10) | 94.9 | 95.9 | 96.2 | 96.3 | 96.5 | 96.7 | 96.9 | 97.2 | 97.3 | 97.3 | 100.0 |

Table 3.11 - LENGTH 5 (Cardinality: 168)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 94.6 | 97.0 | 97.6 | 97.6 | 97.6 | 98.2 | 98.2 | 98.8 | 98.8 | 98.8 | 100.0 |
| 1000 ( 2) | 96.7 | 97.3 | 97.9 | 98.2 | 98.5 | 98.8 | 98.8 | 99.1 | 99.1 | 99.1 | 100.0 |
| 500 ( 2) | 97.0 | 98.2 | 98.8 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 100.0 |
| 200 ( 5) | 98.1 | 99.0 | 99.3 | 99.3 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 100.0 |
| 100 (10) | 98.7 | 99.2 | 99.3 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 99.4 | 100.0 |

Table 3.12 - LENGTH 6 (Cardinality: 157)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 93.6 | 93.6 | 93.6 | 93.6 | 93.6 | 94.3 | 94.3 | 94.3 | 94.3 | 94.3 | 100.0 |
| 1000 ( 2) | 93.6 | 93.6 | 93.6 | 93.9 | 93.9 | 94.6 | 94.9 | 95.5 | 95.5 | 95.5 | 100.0 |
| 500 ( 2) | 93.9 | 94.3 | 94.6 | 94.9 | 95.2 | 95.2 | 95.5 | 95.5 | 95.5 | 95.5 | 100.0 |
| 200 ( 5) | 94.3 | 95.2 | 95.2 | 95.4 | 95.5 | 95.5 | 95.5 | 95.7 | 95.9 | 96.1 | 100.0 |
| 100 (10) | 94.5 | 95.4 | 95.8 | 95.9 | 96.0 | 96.2 | 96.4 | 96.6 | 96.8 | 96.8 | 100.0 |

Table 3.13 - LENGTH 7 (Cardinality: 161)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 94.4 | 95.7 | 95.7 | 95.7 | 95.7 | 95.7 | 95.7 | 96.3 | 96.3 | 96.9 | 100.0 |
| 1000 ( 2) | 95.0 | 95.7 | 95.7 | 96.3 | 96.9 | 96.9 | 96.9 | 96.9 | 96.9 | 96.9 | 100.0 |
| 500 ( 2) | 95.7 | 96.6 | 96.6 | 96.6 | 96.9 | 96.9 | 97.2 | 97.5 | 97.5 | 97.5 | 100.0 |
| 200 ( 5) | 96.0 | 96.4 | 96.8 | 97.1 | 97.4 | 97.8 | 98.1 | 98.5 | 98.5 | 98.6 | 100.0 |
| 100 (10) | 96.6 | 97.3 | 97.5 | 98.1 | 98.6 | 98.7 | 98.7 | 98.8 | 98.8 | 98.8 | 100.0 |

Table 3.14 - LENGTH 8 (Cardinality: 74)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 98.6 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 1000 ( 2) | 98.6 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 500 ( 2) | 99.3 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 200 ( 5) | 99.7 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 100 (10) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Table 3.15 - LENGTH 9 (Cardinality: 51)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 96.1 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 100.0 |
| 1000 ( 2) | 96.1 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 100.0 |
| 500 ( 2) | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 100.0 |
| 200 ( 5) | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 98.0 | 100.0 |
| 100 (10) | 98.0 | 98.0 | 98.0 | 98.0 | 98.4 | 99.2 | 99.4 | 99.6 | 99.6 | 99.6 | 100.0 |

Table 3.16 - LENGTH 10 (Cardinality: 33)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 90.9 | 90.9 | 90.9 | 90.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 100.0 |
| 1000 ( 2) | 90.9 | 92.4 | 92.4 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 100.0 |
| 500 ( 2) | 92.4 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 100.0 |
| 200 ( 5) | 92.7 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 100.0 |
| 100 (10) | 93.3 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 93.9 | 100.0 |

Table 3.17 - LENGTH 11 (Cardinality: 4)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 1000 ( 2) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 500 ( 2) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 200 ( 5) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 100 (10) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Table 3.18 - LENGTH 12 (Cardinality: 2)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 1000 ( 2) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 500 ( 2) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 200 ( 5) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 100 (10) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Table 3.19 - LENGTH 13 (Cardinality: 1)

| Lexicon Size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2000 ( 1) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 1000 ( 2) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 500 ( 2) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 200 ( 5) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 100 (10) | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

By restricting attention to word segmentations which are to the correct number of characters we can estimate an effective character recognition rate for the neural networks. Table 3.20 gives the results of these estimates. The results are given for three cases: 1) all character images vs. all (70) character classes, 2) upper case character images vs. upper case (26) character classes, and 3) lower case character images vs. lower case (26) character classes. For each case, the percent of characters ranked 1, 2 3, and so forth. That is, for characters ranked 1, the top choice of the neural network was the correct character class. For those ranked 2, the correct class was ranked 2, and so forth.

Table 3.20 - Effective Character Recognition Rates for Neural Networks

| DATASET | SIZE | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| All | 5625 | 0.73 | 0.83 | 0.87 | 0.9 | 0.91 | 0.92 | 0.93 |
| Upper Case | 3715 | 0.88 | 0.93 | 0.94 | 0.95 | 0.96 | 0.97 | 0.97 |
| Lower Case | 1901 | 0.87 | 0.92 | 0.93 | 0.94 | 0.94 | 0.95 | 0.96 |

In the word verification test there were a total of 86 images which were not ranked 1 for the largest (2000 string) lexicon. These 86 images are shown in Figure 3.1. The figure shows the strings from the truth file associated with each image, and the string from the lexicon preferred by the word verification module.

| String from Truth File | Top String from Word Verification |
|---|---|
| Realty Road Ro Expressway DR RD Central | Mill Ila FM INGRAM Di El Chelsa |
| TOWN St DR PALADIUM LN Dr RD | MDWS Rt Lk PARADER Of Or Rd |
| ANN Viva RD HILL DRIVE Willow | MAN Usa Rl MILL Delhi Hls |
| Lane Estate BLVD MEADOW BLVD | Cut Cobb 32ND ELDON 32ND |
| MAIN RD CENTRAL Jay Brook | MAIL Rl DUNCAN Av Bronx |
| Maria View RD HILL DR DR | Kamla Wise Rd Hls OR Dl |
| GUS AVE Cr Ct ST RD DR DR DR | FLS RUE Cor CE CT MC Dl OR CR |
| Rd St RD CRCLE Frwy Ferry RD | Fld ST PO Ml Farm Perry Rl |
| Main Timbers DR CLUBHOUSE | Mdw Timber Clf CLIFFDALE |
| COURT Airport DR DR OAK LBJ | CR Annex OR CR CAR UN |
| BEND Addison Rogers Greenville | 32ND Addie MCKAMY Sammy |
| Rorary CIRCLE Mill HILL Quapaw | Peary CRCLE Ml Ml Rpd |
| Cowan Story Wedgewood Freeway | Coach Strm Rd Freewy |
| Highway NOEL Jupiter ROAD | Mckamy Ml Junius Ml |
| Estrada Street 17th Sunny | Leda Jeb LK Ml |
| Lane WOODCASTLE APOLLO | BR MAYES REE |

Figure 3.1 - Word Verification Errors

# 4. END TO END TEST RESULTS

In this section we discuss the testing of the Phase III contextual analysis system. The initial test results are presented in Section 4.1. In Section 4.2 we give a preliminary error analysis which led to the LOS correction process discussed in Section 4.4 and the detailed error analysis discussed in Section 5. In Section 4.3, we discuss the decision strategy used for the test and an alternate decision strategy which was developed in parallel but was not part of the system.

## 4.1. Initial Phase III Test Results

The test results can be summarized as follows. The test dataset consisted of 1013 address block images. Of these images, 399 (39.4%) were assigned correct 9 digit codes, 297 (29.3%) were assigned correct 5 digit codes, 162 (16%) were rejected, and 155 (15.3%) were assigned codes which contained an error. This test was performed using the binary image as input: no information from the truth file or the MLOCR machine was used to arrive at these results.

The results of this test were scored using a "ground truth" set called the LOS (for level of sort) file. Upon examination of the errors, it was found that many of the errors seemed to be due to incorrect or incomplete LOS information. This caused correct results generated by the system to be scored as errors because the result obtained by the system was not among the answers contained in the LOS file. The LOS correction process and revised test results are presented in Section 4.4.

The results of this test are summarized in Table 4.1. The table shows the number of images encoded to each level, and the associated mail processing costs. The table also gives this data for the LOS file results, and the MLOCR BEZ results for comparison.

Table 4.1 - Encode Level and Processing Costs

| Images Encoded to | LOS | LOS Cost | BEZ | BEZ Cost | ERIM | ERIM Cost |
|---|---|---|---|---|---|---|
| 9-digit Direct | 356 | 1.60 | 58 | 0.26 | 136 | 0.61 |
| 9-digit HA | 115 | 1.50 | 21 | 0.27 | 73 | 0.95 |
| 9-digit S | 361 | 7.77 | 67 | 1.44 | 190 | 4.09 |
| 5 digits | 169 | 6.19 | 318 | 11.65 | 272 | 9.96 |
| 3 digits | 10 | 0.45 | 52 | 2.35 | 25 | 1.13 |
| Error Add-on | 0 | 0.00 | 18 | 1.13 | 50 | 3.15 |
| Error 5-digit | 0 | 0.00 | 85 | 7.49 | 105 | 9.25 |
| Unresolved | 2 | 0.10 | 394 | 20.41 | 162 | 8.39 |
| Total | 1013 | 17.62 | 1013 | 45.01 | 1013 | 37.54 |
| Cost Per Thousand | | 17.39 | | 44.43 | | 37.06 |

Table 4.2 shows a confusion matrix where the rows represent ERIM's performance as broken down into categories of sortation and the columns represent the LOS results as broken down by the same categories of sortation.

Table 4.2 - LOS vs ERIM's End to End System

| | 9-Di | 9-HA | 9-H | 9-S | 5 | 3 | Rej | Add | 5-d | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| 9-Di | 136 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 136 |
| 9-HA | 6 | 23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 29 |
| 9-H | 11 | 3 | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 44 |
| 9-S | 29 | 0 | 4 | 157 | 0 | 0 | 0 | 0 | 0 | 190 |
| 5 | 50 | 2 | 21 | 95 | 104 | 0 | 0 | 0 | 0 | 272 |
| 3 | 13 | 0 | 3 | 3 | 5 | 1 | 0 | 0 | 0 | 25 |
| Rej | 68 | 4 | 8 | 48 | 28 | 4 | 2 | 0 | 0 | 162 |
| Add | 13 | 1 | 4 | 13 | 19 | 0 | 0 | 0 | 0 | 50 |
| 5-d | 30 | 6 | 6 | 45 | 13 | 5 | 0 | 0 | 0 | 105 |
| Total | 356 | 39 | 76 | 361 | 169 | 10 | 2 | 0 | 0 | 1013 |

The two columns and rows before the "Total" column/row are error categories where "Add" implies that the base 5-digit ZIP was correct but there was an error in the 4-digit add-on, and "5-d" represents an error in the 5-digit base ZIP Code. In general, positive entries in the upper triangle of this matrix represent improvements over the LOS results and lower triangular numbers indicate lesser performance than the LOS results. Since the LOS results represent an upper bound, this matrix is completely lower triangular.

Similar to Table 4.2, Table 4.3 also shows a confusion matrix where the rows represent levels of sortation achieved by ERIM's system. The two tables differ in the fact that the columns in this new table represent levels of sortation achieved by the MLOCR BEZ as opposed to the LOS results. As contrasted with the previous matrix, this matrix is much denser in its upper triangle than lower, which indicates a significant performance improvement of our contextual analysis system over the MLOCR results.

Table 4.3 - BEZ vs. ERIM's End to End System

| | 9-Di | 9-HA | 9-H | 9-S | 5 | 3 | Rej | Add | 5-d | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| 9-Di | 42 | 0 | 0 | 0 | 29 | 9 | 38 | 2 | 16 | 136 |
| 9-HA | 3 | 8 | 1 | 0 | 9 | 2 | 4 | 0 | 2 | 29 |
| 9-H | 2 | 0 | 8 | 0 | 14 | 3 | 11 | 2 | 4 | 44 |
| 9-S | 0 | 0 | 0 | 46 | 62 | 5 | 55 | 4 | 18 | 190 |
| 5 | 3 | 0 | 3 | 15 | 148 | 12 | 84 | 5 | 2 | 272 |
| 3 | 1 | 0 | 1 | 0 | 3 | 5 | 14 | 0 | 1 | 25 |
| Rej | 2 | 0 | 0 | 0 | 21 | 2 | 130 | 0 | 7 | 162 |
| Add | 0 | 0 | 0 | 1 | 17 | 3 | 22 | 4 | 3 | 50 |
| 5-d | 5 | 0 | 0 | 5 | 15 | 11 | 36 | 1 | 32 | 105 |
| Total | 58 | 8 | 13 | 67 | 318 | 52 | 394 | 18 | 85 | 1013 |

Table 4.4 summarizes the results of the MLOCR (BEZ), ERIM's end to end system, and ERIM's end to end system operating solely on the rejects from the MLOCR. For each of these three cases, the table shows: 1) an encode rate, the percent of images for which the system produces a 9-digit result, 2) an encode error rate, the percent of images encoded which were errors, 3) an accept rate, the percent of images for which the system produced a 5 or 9 digit result, 4) an accept error rate, the percent of accepted images which were in error, and 5) a reject rate, the percent of images rejected by the system.

Table 4.4 - Performance Summaries

| | Total No. of Images | Percent Encode Rate | Percent Error ER | Percent Accept Rate | Percent Error AR | Percent Reject Rate |
|---|---|---|---|---|---|---|
| MLOCR | 1013 | 16.29 | 11.52 | 61.11 | 16.64 | 38.89 |
| ERIM | 1013 | 47.58 | 17.22 | 84.01 | 18.21 | 15.99 |
| ERIM vs. MLOCR Rejects | 394 | 34.52 | 20.59 | 67.01 | 21.97 | 32.99 |

For purposes of comparing the ERIM results with the ECA results, the following four tables are also included. ECA was only able to process 743 of the 1013 images in the Phase III dataset. Tables 4.5 through 4.8 are identical to tables 4.1 through 4.4 except that the results are for 743 images instead of 1013. These 743 images will be referred to as the ECA dataset.

Table 4.5- Encode Level and Processing Costs - ECA Dataset

| Images Encoded to | LOS | LOS Cost | BEZ | BEZ Cost | ERIM | ERIM Cost |
|---|---|---|---|---|---|---|
| 9-digit Direct | 237 | 1.07 | 50 | 0.23 | 108 | 0.49 |
| 9-digit HA | 77 | 1.00 | 20 | 0.26 | 47 | 0.61 |
| 9-digit S | 289 | 6.22 | 66 | 1.42 | 160 | 3.44 |
| 5 digits | 132 | 4.84 | 259 | 9.49 | 222 | 8.13 |
| 3 digits | 7 | 0.31 | 38 | 1.72 | 15 | 0.68 |
| Error Add-on | 0 | 0.00 | 15 | 0.94 | 34 | 2.14 |
| Error 5-digit | 0 | 0.00 | 68 | 5.99 | 75 | 6.61 |
| Unresolved | 1 | 0.05 | 227 | 11.76 | 82 | 4.25 |
| Total | 743 | 13.49 | 743 | 31.81 | 743 | 26.35 |
| Cost Per Thousand | | 18.16 | | 42.81 | | 35.47 |

Table 4.6 - LOS vs ERIM's End to End System - ECA Dataset

| | 9-Di | 9-HA | 9-H | 9-S | 5 | 3 | Rej | Add | 5-d | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| 9-Di | 108 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 108 |
| 9-HA | 4 | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 |
| 9-H | 6 | 2 | 21 | 0 | 0 | 0 | 0 | 0 | 0 | 29 |
| 9-S | 26 | 0 | 3 | 131 | 0 | 0 | 0 | 0 | 0 | 160 |
| 5 | 31 | 2 | 16 | 77 | 96 | 0 | 0 | 0 | 0 | 222 |
| 3 | 7 | 0 | 3 | 2 | 2 | 1 | 0 | 0 | 0 | 15 |
| Rej | 31 | 1 | 4 | 31 | 13 | 1 | 1 | 0 | 0 | 82 |
| Add | 7 | 1 | 4 | 10 | 12 | 0 | 0 | 0 | 0 | 34 |
| 5-d | 17 | 4 | 2 | 38 | 9 | 5 | 0 | 0 | 0 | 75 |
| Total | 237 | 24 | 53 | 289 | 132 | 7 | 1 | 0 | 0 | 743 |

Table 4.7 - BEZ vs. ERIM's End to End System - ECA Dataset

| | 9-Di | 9-HA | 9-H | 9-S | 5 | 3 | Rej | Add | 5-d | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| 9-Di | 39 | 0 | 0 | 0 | 21 | 6 | 24 | 2 | 16 | 108 |
| 9-HA | 3 | 8 | 0 | 0 | 4 | 1 | 1 | 0 | 1 | 18 |
| 9-H | 1 | 0 | 8 | 0 | 10 | 2 | 4 | 2 | 2 | 29 |
| 9-S | 0 | 0 | 0 | 46 | 52 | 4 | 40 | 4 | 14 | 160 |
| 5 | 2 | 0 | 3 | 14 | 130 | 10 | 56 | 5 | 2 | 222 |
| 3 | 1 | 0 | 1 | 0 | 2 | 3 | 7 | 0 | 1 | 15 |
| Rej | 2 | 0 | 0 | 0 | 16 | 1 | 58 | 0 | 5 | 82 |
| Add | 0 | 0 | 0 | 1 | 12 | 2 | 14 | 2 | 3 | 34 |
| 5-d | 2 | 0 | 0 | 5 | 12 | 9 | 23 | 0 | 24 | 75 |
| Total | 50 | 8 | 12 | 66 | 259 | 38 | 227 | 15 | 68 | 743 |

Table 4.8 - Performance Summaries - ECA Dataset

| | Total No. of Images | Percent Encode Rate | Percent Error ER | Percent Accept Rate | Percent Error AR | Percent Reject Rate |
|---|---|---|---|---|---|---|
| MLOCR | 743 | 20.32 | 9.93 | 69.45 | 16.09 | 30.55 |
| ERIM | 743 | 50.34 | 15.77 | 88.96 | 16.49 | 11.04 |
| ERIM vs. MLOCR Rejects | 227 | 38.77 | 21.59 | 74.45 | 21.89 | 25.55 |

## 4.2. Preliminary Error Analysis

According to the LOS information, ERIM's end to end system came up with a total of 155 errors. A rough breakdown of the causes of these errors is shown in Table 4.9 below. The largest category is false errors. Our review of the results shows that 68 of the errors that the system is charged with are not really errors. There are variety of reasons for this, which will be described in detail below. The next most common cause of errors are patron errors in the city and/or ZIP code. The Phase III test set consisted of a large number of images in which the city and ZIP code agree with each other, but the ZIP code is incorrect. In these cases, if the Contextual Analysis portion of the system fails to come up with a nine-digit result, there is no way to avoid an error. The next most frequent problem is numeral reading, which caused 21 errors. Bad word verifications of city/ZIP words used by fuzzy logic city/ZIP reader caused another five errors. Incompleteness of the NCWS information caused four errors. This refers to the situation when the primary number on the mailpiece is not among the primary numbers listed in the NCWS for that street, but other primary numbers are. It is presumed that this same problem was the cause of

some number of system rejects as well. Finally, the remaining 22 errors are caused by other miscellaneous problems.

Table 4.9 - Breakdown of System Errors

| Error Type | Number |
|---|---|
| False Errors | 68 |
| Patron Errors | 25 |
| Numeral Reading | 21 |
| Bad Word Verification of ZIP/City | 5 |
| Incomplete NCWS Information | 4 |
| Other | 22 |
| Total | 145 |

As stated above, we believe that 68 of the system errors are not really errors. The discrepancies for the most part fall into four categories as shown in Table 4.10.

Table 4.10 - Breakdown of False Errors

| Error Type | Number |
|---|---|
| ZIP Discrepancy | 18 |
| Default City ZIP | 20 |
| 5-Digit Direct ZIP | 7 |
| NCWS Information | 4 |
| Other | 19 |
| Total | 68 |

The first category, which is the source of 18 of the false errors, is the ZIP discrepancy problem. In each of these cases the LOS record gives a response in one ZIP code, and we feel that the correct nine-digit ZIP is in a different ZIP code. There seems to be a difference in the ZIP codes which the two databases have. In many cases the LOS record contains a result in 75093, a ZIP code which is in our city/state file, but for which we have no ZIP+4 records. It should be noted that many times the add-on portions of the two ZIPs match. This suggests that there may have been a rezoning which took place. Also, of the 18 images, 13 times the mailpiece agrees with ERIM's ZIP, and 5 times the mailpiece agrees with the LOS ZIP.

The second cause of false errors are LOS errors in the default city ZIP for a record. This problem occurred 20 times. Typically, instead of encoding the true city default ZIP, the LOS record just backs off to a double-zero ZIP, e.g. 75100.

The third problem concerns the treatment of 5 digit direct ZIPs. When a 5-digit direct ZIP is found on a mailpiece, rather than proceed with a contextual analysis of the mailpiece, the mailpiece is sent directly to that ZIP, and a 9 digit direct cost is charged. Seven times, the LOS record ignores this rule and performs a contextual analysis, coming up with a different result. Also, it should be noted that there are an additional 15 cases where the ERIM system assigns a 5 digit direct ZIP to a mailpiece, and the mailpiece is scored as a 5 digit code, with cost 36.63 rather than the correct cost of 4.50.

The fourth major cause of false errors concerns the use of NCWS information. In some cases, NCWS information can be used to disambiguate between two records in the ZIP+4 database. For example, consider a mailpiece to 108 Main Street. If, in the database, there is a record for 100-198 North Main Street and a record for 100-198 South Main Street, there is no way to know which one is meant, so the most specific correct ZIP is to 5 digits. However, the corresponding NCWS records might indicate that there is 108 North Main Street but no 108 South Main Street. In this case it is appropriate to assign the 9 digit code for North Main Street. This happened four times in the Phase III test set. The 19 remaining false errors are caused by other reasons. In general, these consist of typographical/clerical errors (for example transposed digits in the ZIP code) or an insufficient search for close street names in the database (for example missing a database record for FIELD TRAIL DR for a mailpiece which omitted the suffix making TRAIL appear to be the suffix).

Table 4.11 gives breakdown of the 1013 images according to which component of the system decision module responded with the end-to-end answer and the number of errors attributed to each component.

Table 4.11 - Breakdown of Responses by System Components

| Component | No. Responses | No. Errors | No. True Errors |
|---|---|---|---|
| 5DD Numeral Reader | 32 | 10 | 3 |
| 5DD Fuzzy Logic | 3 | 1 | 1 |
| PO Box System | 142 | 20 | 13 |
| Contextual Analysis | 350 | 67 | 33 |
| Fuzzy Logic 5 or 3 | 228 | 36 | 29 |
| Numeral Read 9 | 6 | 1 | 1 |
| Numeral Read 5 | 54 | 9 | 7 |
| Numeral Read 3 | 36 | 11 | 0 |
| Reject | 162 | 0 | 0 |
| Total | 1013 | 155 | 87 |

## 4.3. Decision Strategies

One area which received attention during the latter portions of Phase III is error reduction. In an effort to reduce errors, we introduced the notion of address block verification, an attempt to go back and match a candidate ZIP+4 record to the mailpiece image. Along with address block verification, a decision strategy is needed to decide when a candidate ZIP+4 record matches well enough to respond with the corresponding ZIP+4 code. It needs to use the cost model to be able to choose the record with the lowest expected cost among many potentially correct records. In using a decision strategy, there will always be some number of correct responses which are rejected because they overlap with incorrect responses in the space of confidence of match. To get a better feel for why the system rejects an image, it is good to know whether it is due to the contextual analysis system being entirely off the track, or whether it is due to the confidence resulting from the match being low. The results of the Phase III test show that 399 images are assigned a correct ZIP+4 code. An analysis of the results showed that in 474 of the images, a correct ZIP+4 code is found. Thus, 75 potentially correct images are rejected in an effort to keep the error rate low.

In the time prior to the Phase III test, two decision strategies were being developed, one which used Gaussian curves of correct and incorrect record-to-image matches, and one which used the Dempster-Shafer theory of evidential reasoning. Because there was insufficient time to fully develop the Dempster-Shafer decision strategy, and because the Gaussian strategy performed better on a test set of 929 images, the Gaussian strategy was used for the Phase III test. Table 4.12 below compares the results using the Gaussian strategy and the Dempster-Shafer strategy. As can be seen from the table, the Dempster-Shafer strategy performs better on this set of images, assigning 39 more correct 9 digit ZIPs at an expense of 9 errors, with the total cost decreasing by 0.67. Overall, when considered in isolation from the rest of the system, the Dempster-Shafer strategy does better than the Gaussian on 44 images, and the Gaussian does better than the Dempster-Shafer on 29 images. This implies that there is significant room for improvement in both strategies. One of the areas we anticipate spending time on during Phase IV is the continued development of these decision strategies leading to the development of a strategy which takes advantage of the best features of each.

Table 4.12 - Comparison of the Two Decision Strategies

| Images Encoded to | Gaussian Strategy | Gaussian Cost | Dempster-Shafer Strategy | Dempster-Shafer Cost |
|---|---|---|---|---|
| 9-digit Direct | 136 | 0.61 | 136 | 0.61 |
| 9-digit HA | 73 | 0.95 | 88 | 1.15 |
| 9-digit S | 190 | 4.09 | 214 | 4.61 |
| 5 digits | 272 | 9.96 | 236 | 8.64 |
| 3 digits | 25 | 1.13 | 24 | 1.09 |
| Error Add-on | 50 | 3.15 | 60 | 3.78 |
| Error 5-digit | 105 | 9.25 | 104 | 9.17 |
| Unresolved | 162 | 8.39 | 151 | 7.82 |
| Total | 1013 | 37.54 | 1013 | 36.87 |
| Cost Per Thousand | | 37.06 | | 36.40 |

## 4.4. LOS File Correction and Revised Phase III Test Results

This section describes the development of a true LOS file for the 1013 images from the Phase III test. As noted in Section 4.2, the original LOS file contains many errors, in part due to a discrepancy between the database used to generate the LOS file, and that used in the contextual analysis system. By developing a true LOS file, it is possible to more accurately assess overall system performance. Also, we can gain insight about why so many images are at best encoded to 5 digits.

In creating the true LOS file, several policy issues about where to send mailpieces arose. For each issue, we have adopted a consistent attitude about scoring. For images with both a street address and a post office box address, both addresses were scored as correct. For images where the street name cannot be found in either the city or ZIP on the mailpiece, it is not proper to back off this information and send to a different city and ZIP in the four SCFs. For images where the the city and ZIP disagree with each other, and the street name does not exist in either the city or the ZIP, the proper destination is to the ZIP. (A more conservative thing to do would be to reject the mailpiece.) For images with a 5 digit direct ZIP, the mailpiece is sent to that ZIP, regardless of what might appear elsewhere on the image. For images with a 9 digit ZIP on the mailpiece, that ZIP is only correct if a corresponding ZIP+4 record for the ZIP exists, and only if that record matches what is present in the image.

The true LOS file was created as follows. First, for each 9 digit ZIP on each original LOS record, the corresponding database record(s) were retrieved. In 43 cases this process failed, i.e. in our database there is no record corresponding to the given ZIP. These images were reviewed manually and a revised truth was assigned. Next, 59 images which were flagged as false errors

in the original analysis were reviewed manually and assigned a revised truth. From the remaining images, all those which our system did not successfully assign a 9 digit ZIP to were examined manually, during the process of error and reject analysis. This set necessarily included all images for which the original LOS record encoded to only 5 digits (LOS-fives). During this review, 50 images had their truth changed. Thus, in total 152 images had their truth changed.

The most difficult images to review were those for which the LOS had only a 5 digit response. To try to find a match, a variety of querying techniques were used. These queries included ZIP, city, suffix, directional, and primary number queries, as well as queries about primary words. In several cases, a brute force attempt was made, manually examining all primary names in a given ZIP code. For images without a primary line, all secondary names in the given ZIP code were manually examined. While it is not absolutely certain that all LOS-fives which can potentially be improved were in fact improved, we feel that our analysis came close.

Having created a true LOS file, it is possible to rescore the Phase III test run. Table 4.13 summarizes the improvement with respect to encode level. Tables 4.14-4.17 are modified versions of Tables 4.1-4.4. Note that the changes to the LOS have been reflected in the LOS and BEZ scores as well.

Table 4.13 - LOS Correction Summary

| Images Encoded to | Before | After | Change |
|---|---|---|---|
| 9-digit Direct | 356 | 370 | +14 |
| 9-digit HA | 115 | 118 | +3 |
| 9-digit S | 361 | 381 | +20 |
| 5 digits | 169 | 138 | -31 |
| 3 digits | 10 | 4 | -6 |
| Error Add-on | 0 | 0 | 0 |
| Error 5-digit | 0 | 0 | 0 |
| Unresolved | 2 | 2 | 0 |
| Total | 1013 | 1013 | 37 |

Table 4.14 - Encode Level and Processing Costs - Corrected LOS

| Images Encoded to | LOS | LOS Cost | BEZ | BEZ Cost | ERIM | ERIM Cost |
|---|---|---|---|---|---|---|
| 9-digit Direct | 370 | 1.67 | 62 | 0.28 | 165 | 0.74 |
| 9-digit HA | 118 | 1.54 | 21 | 0.27 | 79 | 1.03 |
| 9-digit S | 381 | 8.20 | 63 | 1.36 | 215 | 4.63 |
| 5 digits | 138 | 5.05 | 331 | 12.12 | 270 | 9.89 |
| 3 digits | 4 | 0.18 | 46 | 2.08 | 43 | 1.95 |
| Error Add-on | 0 | 0.00 | 18 | 1.13 | 30 | 1.89 |
| Error 5-digit | 0 | 0.00 | 78 | 6.87 | 49 | 4.32 |
| Unresolved | 2 | 0.10 | 394 | 20.41 | 162 | 8.39 |
| Total | 1013 | 16.74 | 1013 | 44.53 | 1013 | 32.83 |
| Cost Per Thousand | | 16.53 | | 43.96 | | 32.41 |

Table 4.15 - LOS vs. ERIM's End to End System - Corrected LOS

| | 9-Di | 9-HA | 9-H | 9-S | 5 | 3 | Rej | Add | 5-d | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| 9-Di | 165 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 165 |
| 9-HA | 6 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 31 |
| 9-H | 11 | 4 | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 48 |
| 9-S | 28 | 0 | 4 | 183 | 0 | 0 | 0 | 0 | 0 | 215 |
| 5 | 46 | 3 | 23 | 106 | 92 | 0 | 0 | 0 | 0 | 270 |
| 3 | 20 | 1 | 3 | 13 | 5 | 1 | 0 | 0 | 0 | 43 |
| Rej | 72 | 4 | 11 | 51 | 19 | 3 | 2 | 0 | 0 | 162 |
| Add | 8 | 1 | 3 | 11 | 7 | 0 | 0 | 0 | 0 | 30 |
| 5-d | 14 | 2 | 1 | 17 | 15 | 0 | 0 | 0 | 0 | 49 |
| Total | 370 | 40 | 78 | 381 | 138 | 4 | 2 | 0 | 0 | 0 |

Table 4.16 - BEZ vs. ERIM's End to End System - Corrected LOS

| | 9-Di | 9-HA | 9-H | 9-S | 5 | 3 | Rej | Add | 5-d | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| 9-Di | 48 | 0 | 0 | 0 | 30 | 14 | 47 | 4 | 22 | 165 |
| 9-HA | 3 | 8 | 1 | 0 | 10 | 2 | 5 | 0 | 2 | 31 |
| 9-H | 2 | 0 | 8 | 0 | 17 | 3 | 12 | 2 | 4 | 48 |
| 9-S | 0 | 0 | 0 | 46 | 75 | 7 | 60 | 4 | 23 | 215 |
| 5 | 3 | 0 | 3 | 16 | 150 | 9 | 83 | 3 | 3 | 270 |
| 3 | 1 | 0 | 1 | 0 | 5 | 6 | 25 | 1 | 4 | 43 |
| Rej | 2 | 0 | 0 | 0 | 22 | 2 | 130 | 0 | 6 | 162 |
| Add | 0 | 0 | 0 | 0 | 9 | 1 | 16 | 2 | 2 | 30 |
| 5-d | 3 | 0 | 0 | 1 | 13 | 2 | 16 | 2 | 12 | 49 |
| Total | 62 | 8 | 13 | 63 | 331 | 46 | 394 | 18 | 78 | 0 |

Table 4.17 - Performance Summaries - Corrected LOS

| | Total No. of Images | Percent Encode Rate | Percent Error ER | Percent Accept Rate | Percent Error AR | Percent Reject Rate |
|---|---|---|---|---|---|---|
| MLOCR | 1013 | 16.68 | 13.61 | 61.11 | 15.51 | 38.89 |
| ERIM | 1013 | 50.05 | 9.47 | 84.01 | 9.28 | 15.99 |
| ERIM vs. MLOCR Rejects | 394 | 36.80 | 14.48 | 67.01 | 12.12 | 32.99 |

# 5. SYSTEM ERROR ANALYSIS

## 5.1. Overview

Of the 1013 address blocks in the Phase III test set the system was able to correctly (after LOS correction) assign 9-digit ZIPs to 459 mailpieces leaving 554 mailpieces with incorrect or not to 9-digit assignments. Figure 5.1 shows the number of mailpieces for each encode level. This figure corresponds to the ERIM column in Table 4.14. Figure 5.2 shows the same information in a pie chart.

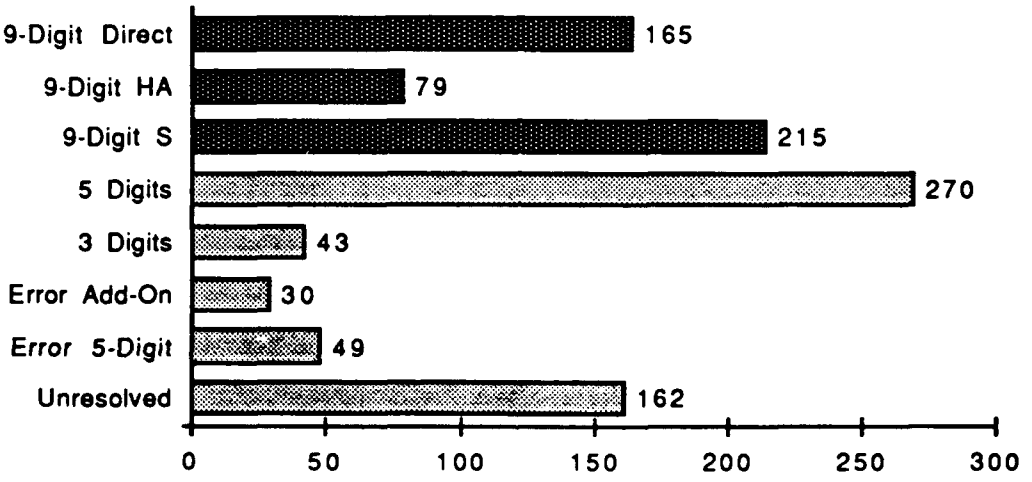



Figure 5.1 - Encode level statistics for Phase III test.

Our error analysis concentrates on the 554 problem mailpieces. The problems with these mailpieces fall into four categories: LOS-5, Image Processing, Matching, and Both Image Processing and Matching. In Figure 5.3, the number of mailpieces in each of these categories (Total Images) is shown along with the number of problems (Total Errors). There are more problems than mailpieces because some mailpieces suffer from multiple problems. The same information is shown in pie charts in Figures 5.4 and 5.5. Figure 5.3 shows that the general category of image processing errors is the largest group of errors. The next largest category, labeled LOS-5, includes those images for which a 9-digit ZIP code was not obtainable through a manual exploration of the database. The size of this category motivates the use of business names and surnames for the address recognition process. The next largest category of images includes both image processing and matching errors. That is, not only did the image processing have a problem, but the matching process employed in the Phase III system was judged to be inadequate for these cases. The last group includes images which failed purely for matching issues. Clearly, some improvements in the matching strategy will be necessary. The categories are further subdivided and discussed in detail in the following sections.

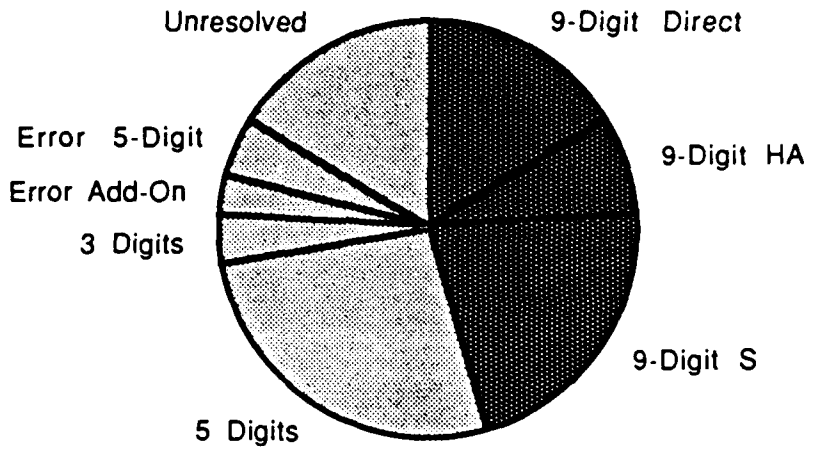



Figure 5.2 - Pie chart of encode level statistics for Phase III test.

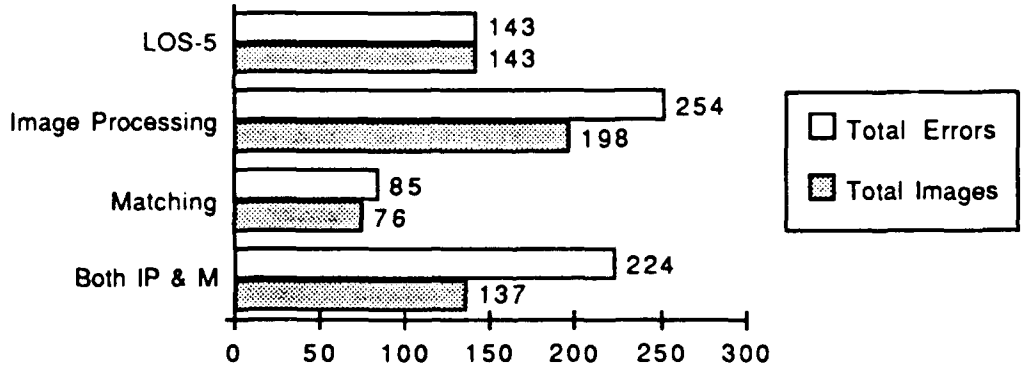



Figure 5.3 - Error count and image count breakdown.

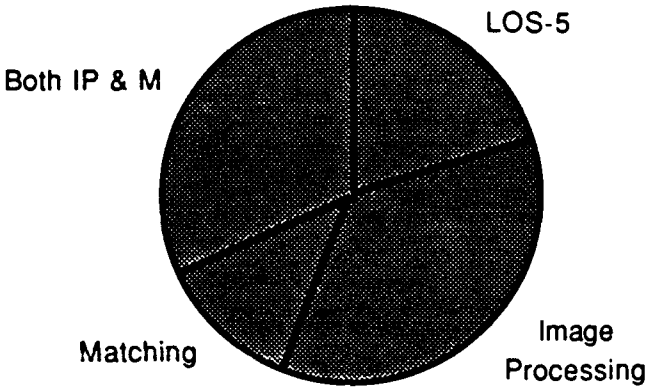



Figure 5.4 - Error count breakdown.

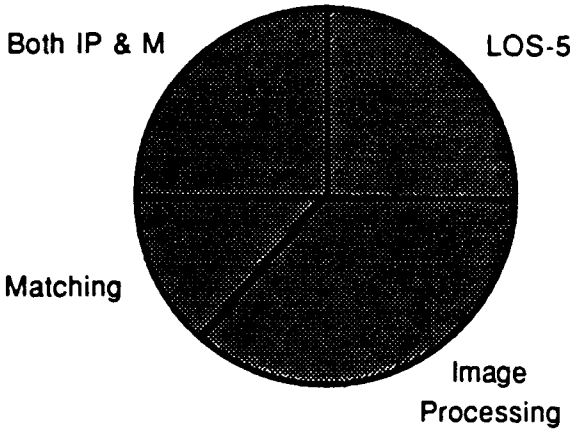



Figure 5.5- Image count breakdown.

## 5.2. LOS-5

The LOS-5 category refers to mailpieces for which no 9-digit ZIP can be determined. As shown in Figure 5.6 there are several reasons for this problem.

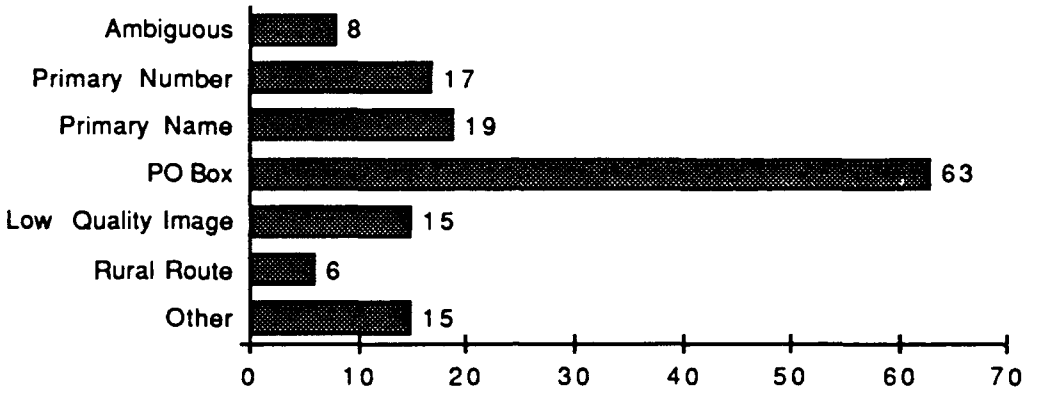



Figure 5.6 - Breakdown of LOS-5 problems.

### Ambiguous

A mailpiece is ambiguous when there is insufficient information on the mailpiece to determine which of several potential ZIP+4 records is the correct one. In the example shown in Figure 5.7 below, the pre-directional has been omitted from the mailpiece. As a result, it is not possible to decide whether the proper destination is to North Shiloh Road or South Shiloh Road. The best that can be done is to send the mailpiece to ZIP 75042.

**High Quality Widgets , Inc.**
Battery Division
1111 Shiloh Road
Garland. TX 75040

| Pnum Range | Pre-dir | Primary | Suffix | City | ZIP |
|---|---|---|---|---|---|
| 1101 1199 O | N | Shiloh | Rd | Garland | 75042-5723 |
| 1102 1199 O | S | Shiloh | Rd | Garland | 75042-8047 |

Figure 5.7- An ambiguous LOS-5 mailpiece and two corresponding ZIP+4 records.

### Primary Number

The primary number problem occurs when database records exist for the primary name on the mailpiece in either the city or ZIP specified, but none of the range records for that primary name match the primary number on the mailpiece. Figure 5.8 below shows an example of this problem. In the database, the legal primary numbers for "Arborcrest" range from 300 to 499. Primary number problems are probably a result of either incomplete database records or patron errors in the primary number.

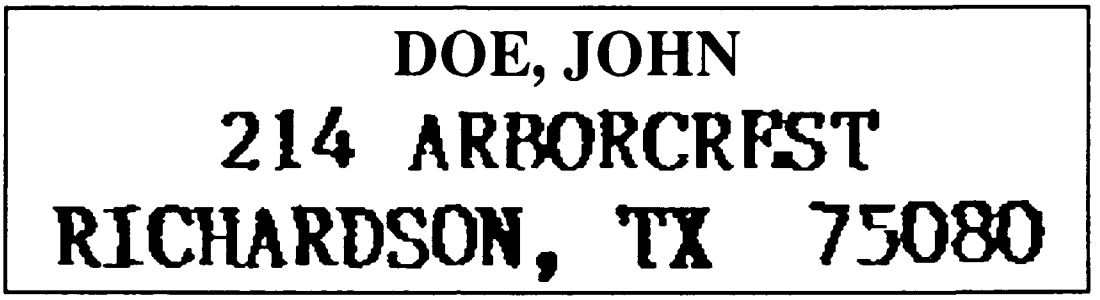



Figure 5.8 - An example of the LOS-5 primary number problem.

### Primary Name

The primary name problem occurs when the primary name on the mailpiece does not exist in either the city or the ZIP on the mailpiece. Figure 5.9 shows an example of this. In our database, the primary name "Shady Oak" occurs twice - once in Dallas in ZIP 75229, and once in Grand Prairie in ZIP 75052. Since neither of these correspond to the localities described on the mailpiece, the image is categorized as an LOS-5 primary name problem.

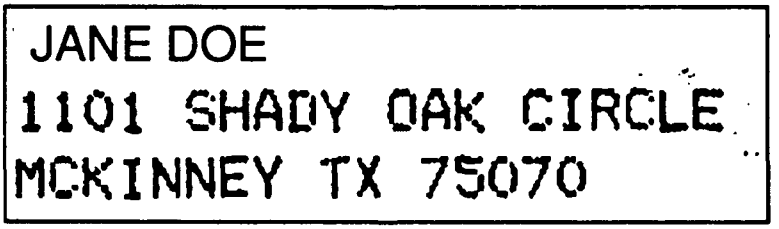



Figure 5.9 - An example of the LOS-5 primary name problem.

### PO Box

The PO Box problem occurs when the PO Box number on the mailpiece does not correspond to any PO Box record in the city or ZIP specified on the mailpiece. Figure 5.10 below shows an example of this. Since the only ZIP in "The Colony" is 75056, the only records which need to be considered are in that ZIP. For all PO Box records in 75056, the first two digits are 56. Thus, no match exists.

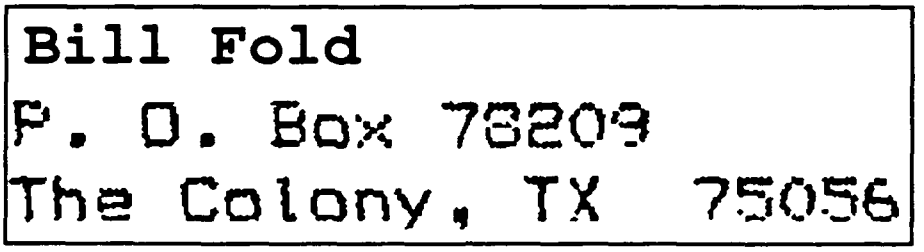



Figure 5.10 - An example of the LOS-5 PO Box problem.

### Low Quality Image

Low quality images are images where the text of the mailpiece is unreadable due to poor image quality. An example of this is shown in Figure 5.11 below.

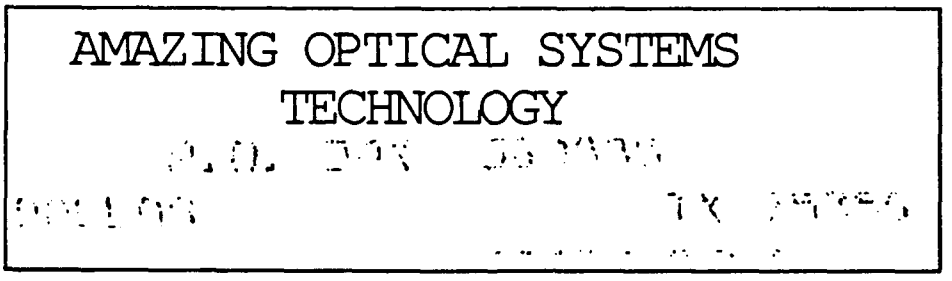



Figure 5.11 - An example of the LOS-5 low quality image problem.

### Rural Route

The rural route problem occurs when the route number or route box number on the mailpiece does not correspond to any rural route record in the city or ZIP specified on the mailpiece. Figure 5.12 below shows an example of this. No rural route records exist for ZIP 75010. In Carrollton there are two rural routes, but both are route 1.

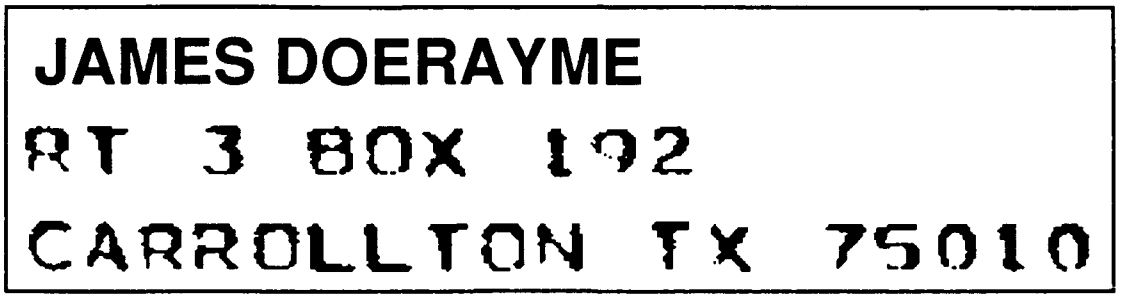
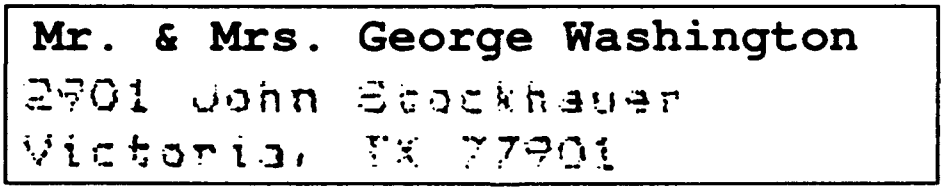
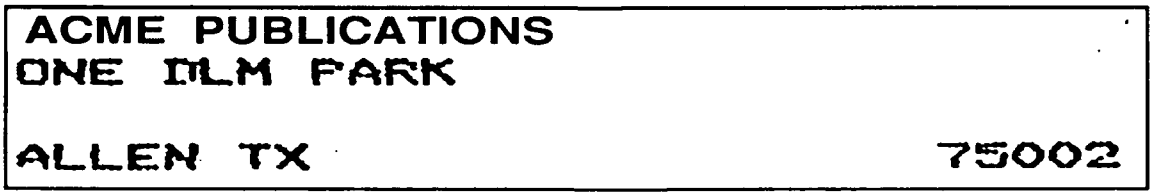



Figure 5.12 - An example of the LOS-5 rural route problem.

**Other**

The remainder of LOS-5 images are caused by a variety of different problems. Three examples are shown below in Figure 5.13. In the first, the address falls outside of the four SCFs under consideration. In the second, there is no primary information on the mailpiece, and the secondary information does not match anything in the database. In the third example, there is also no primary information on the mailpiece.



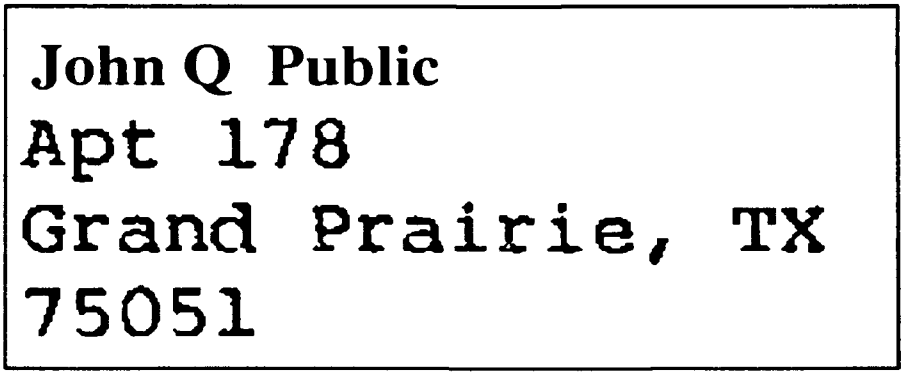




Figure 5.13 - LOS-5s caused by other problems.

## 5.3. Image Processing

Figure 5.14 shows the number of times each of the image processing error types were observed.

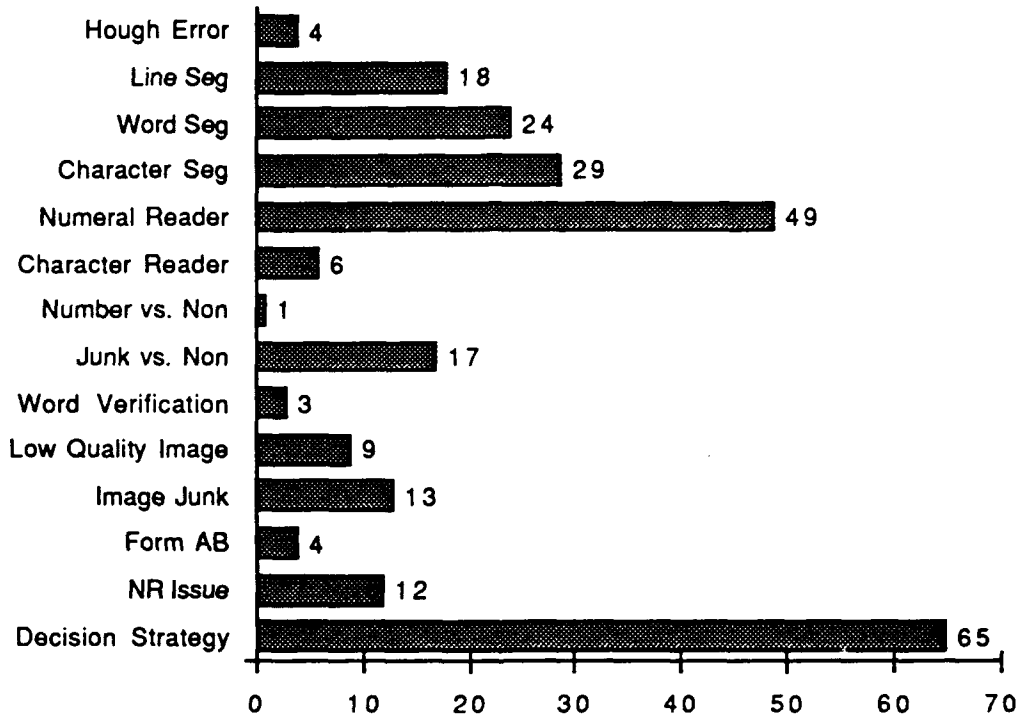



Figure 5.14 - Breakdown of image processing errors.

### Hough Error

The failure of the Hough transform process for de-skewing results in a Hough error. An example is shown in Figure 5.15 below.

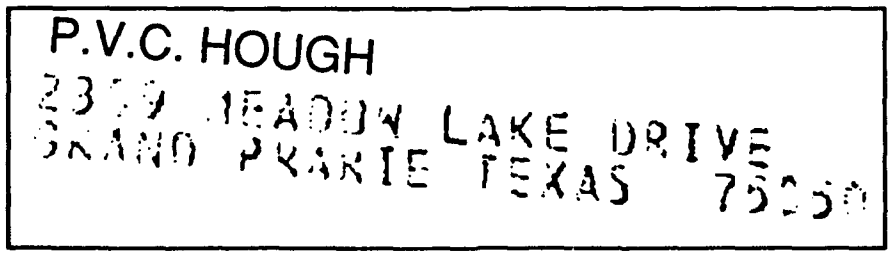



Figure 5.15 - A Hough Error.

### Line Segmentation

Line segmentation errors occur when the address block is not correctly segmented into separate lines. An example is shown in Figure 5.16 below. In this case, the line segmenter has trimmed off the tops of the characters of the first line.

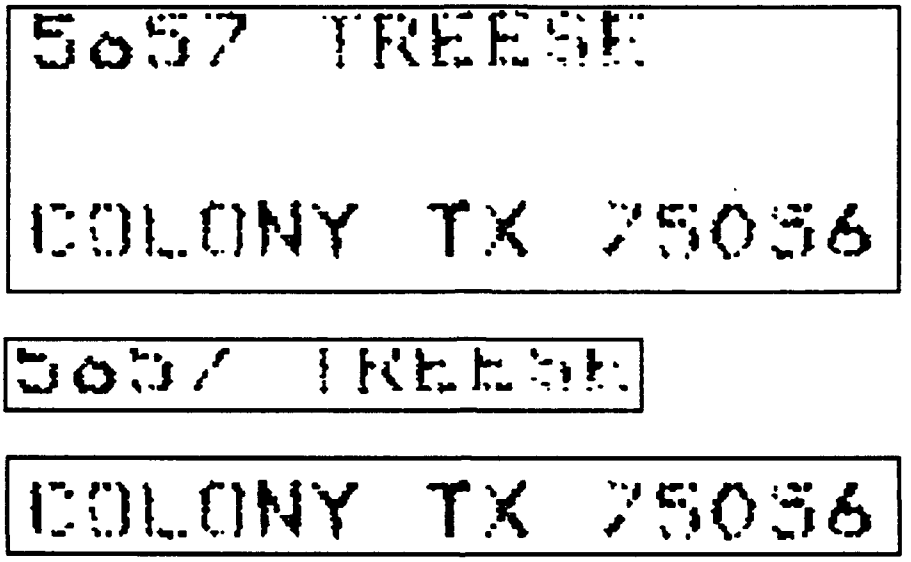


Figure 5.16 - A line segmentation error.

### Word Segmentation

Word segmentation errors occur when a line is not correctly segmented into words. Figure 5.17 below shows an example of this. In this case, none the of the word segmentation hypotheses correctly split the box and box number fields into two words.

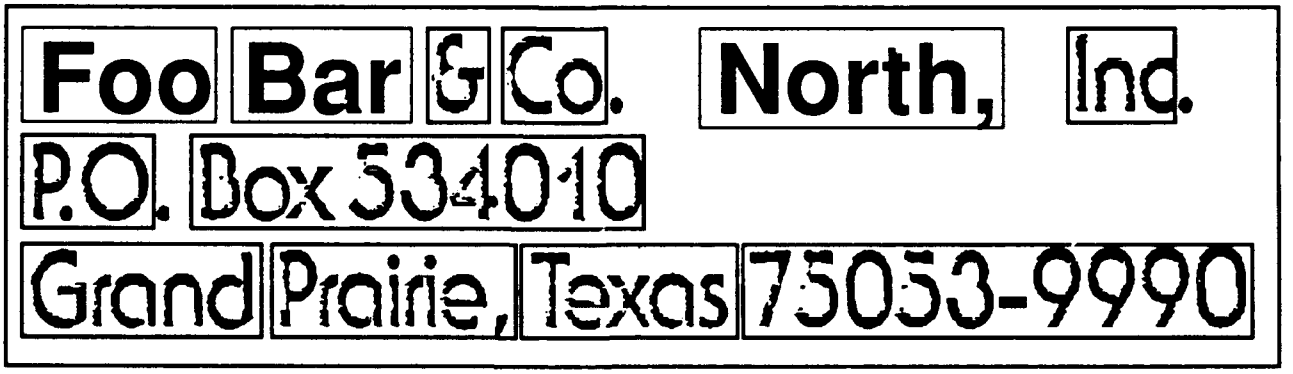



Figure 5.17 - A word segmentation error.

### Character Segmentation

Character segmentation errors occur when a word is not correctly segmented into characters. See Figure 5.18 below.

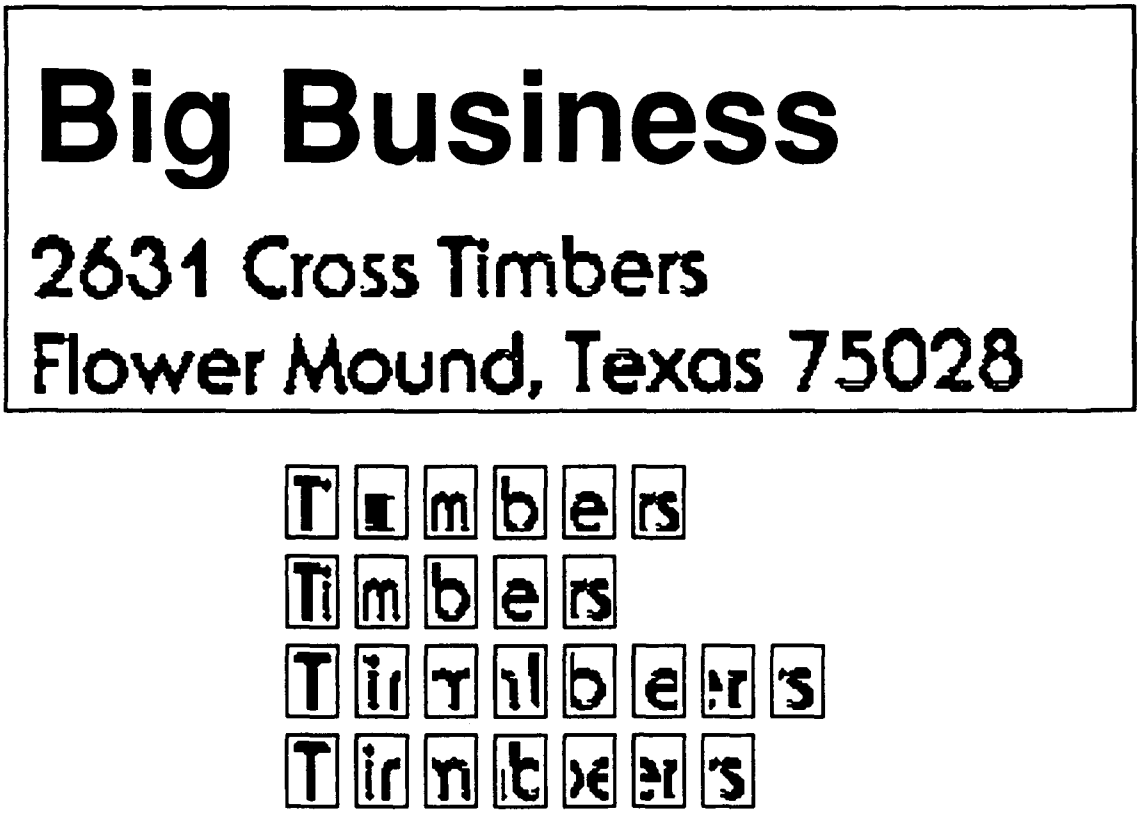



Figure 5.18 - A character segmentation error.

### Numeral Reader

Numeral reading errors occur when the numeral reader reads a number incorrectly. In the example in Figure 5.19 below, the PO Box number is incorrectly read as 330026.

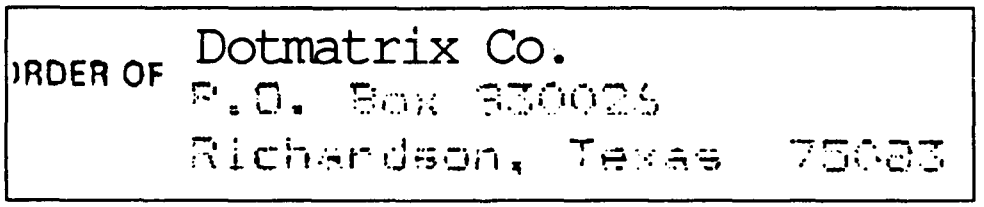



Figure 5.19 - A numeral reading error.

### Character Reader

Character reading errors occur when the character reader fails to correctly read 2 or more out of 4 characters at the start and end of the primary word. When character recognition fails, the correct primary word is pruned from the lexicon and never verified. In the example of Figure 5.20, "MAIN" has been read as "XALT." Since only one character is correct, "MAIN" is pruned from the lexicon and not verified.

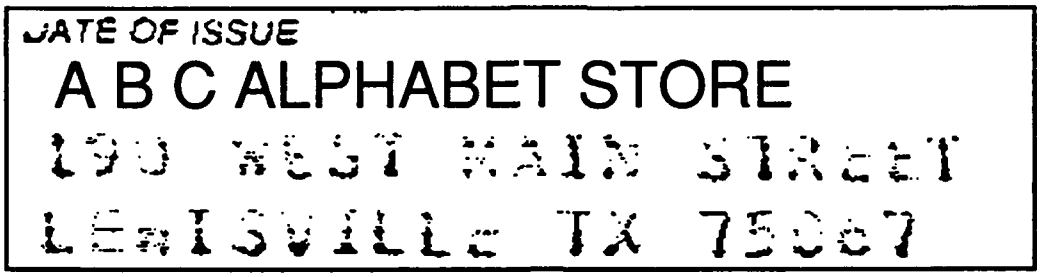
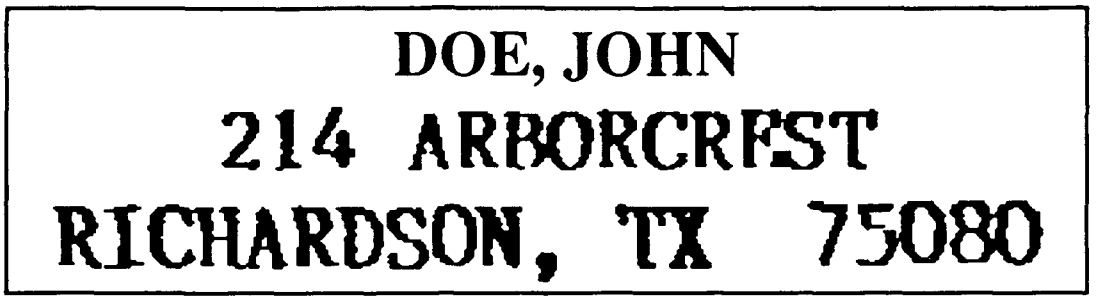



Figure 5.20 - A character reading error.

### Junk vs. Non

Junk vs non errors occur when the neural net ranks incorrect character segmentations higher than the correct segmentation. Figure 5.21below shows the character segmentations for an example primary word along with the associated junk vs. non ratings.



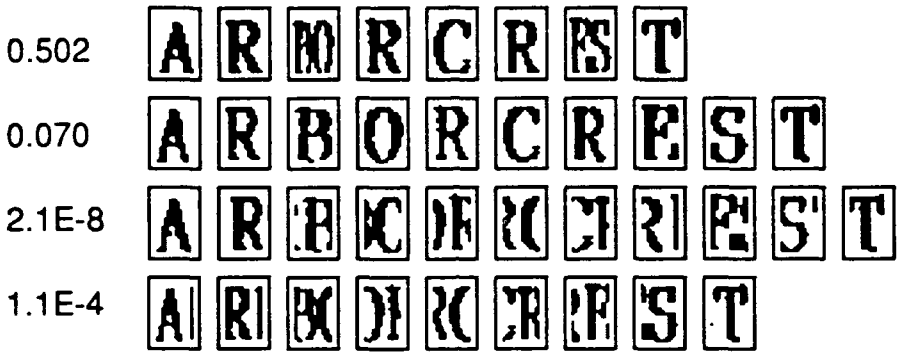




Figure 5.21 - A junk vs. non error.

**Number vs. Non**

Number vs non errors occur when the neural net strongly rates a number as a non-number or strongly rates a non-number as a number. In Figure 5.22 below, the number vs. non score for the primary number is low. Since this value is multiplied into the score which rates the match of the primary number to a primary range, a low match confidence results and the mailpiece is rejected.

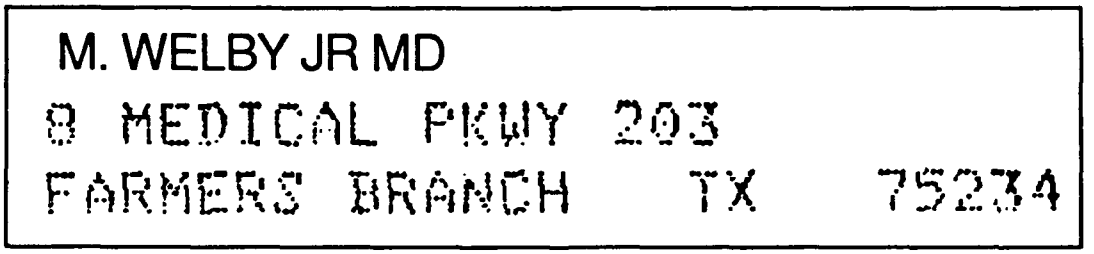



Figure 5.22 - A number vs. non error.

### Word Verification

Word verification errors occur when the word verification process ranks an incorrect primary word higher than the true primary word. In Figure 5.23 below, "SUDAN" (0.102) is verified higher than "SHOAF" (0.096).

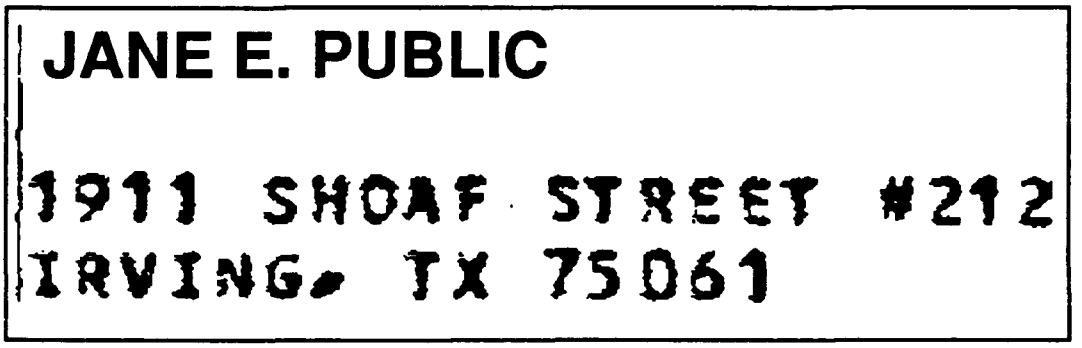



Figure 5.23 - A word verification error.

### Low Quality Image

Low quality image errors occur when one or more system errors are caused by exceptionally poor image quality. Figure 5.24 shows one such image.

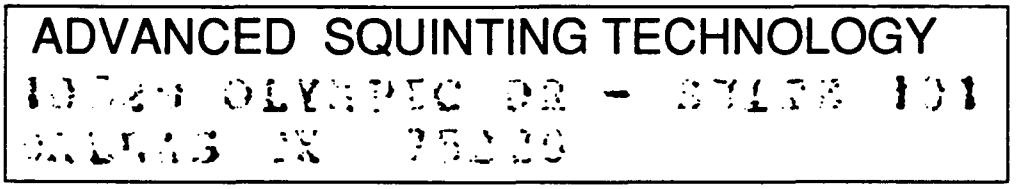



Figure 5.24 - A low quality image error.

### Image Junk

Image junk errors occur when extraneous junk on the mailpiece confuses the system. Among the causes of image junk errors are: speckle caused by plastic windowed envelopes, horizontal or vertical lines passing though or near the image, stray pen or pencil marks, and logos. Figure 5.25 below shows some examples.

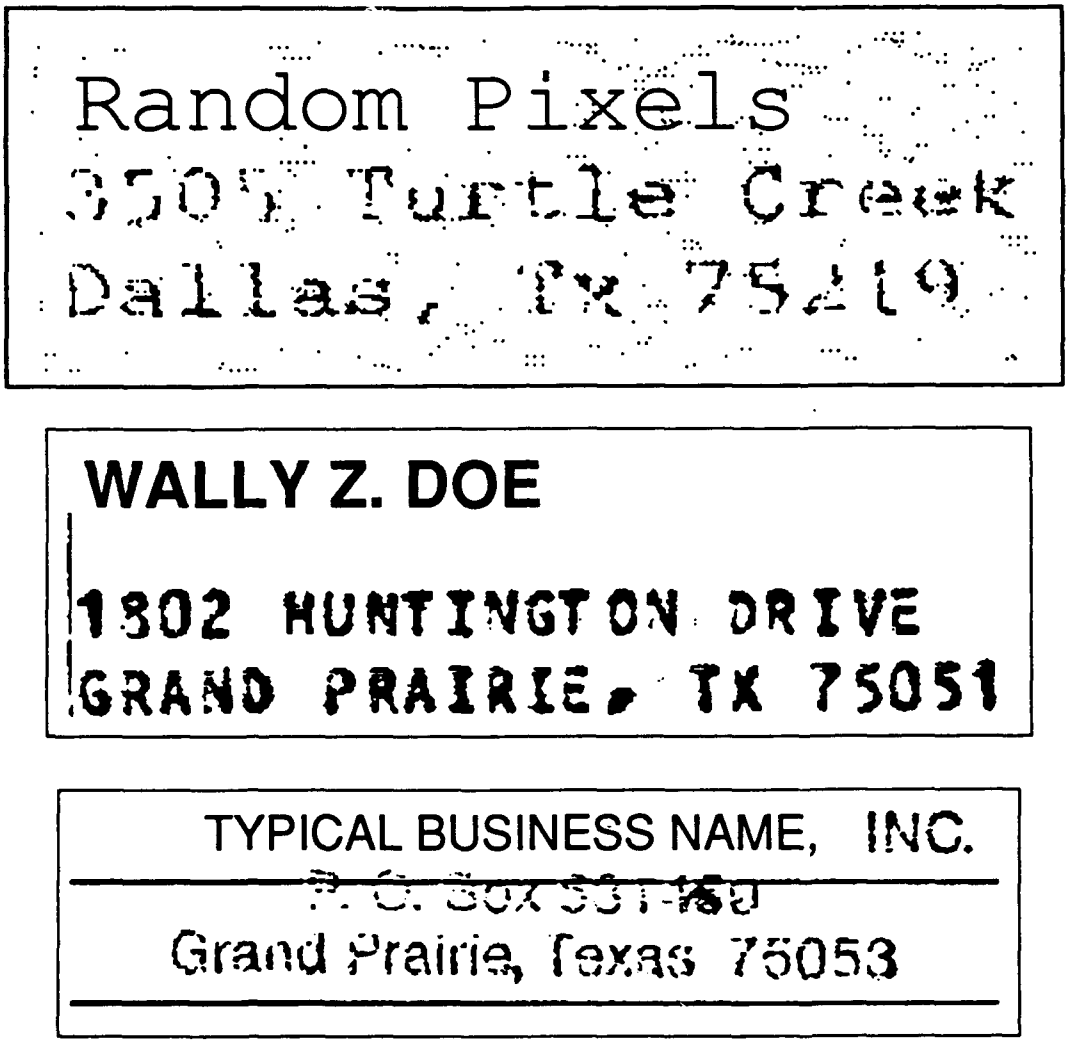



Figure 5.25 - Some errors caused by image junk.

**Form AB**

Form address block errors can be thought of as a special case of image junk. They occur when the address block is superimposed on a form which has underlines or labels directing the patron where to write the address. An example is shown in Figure 5.26 below.

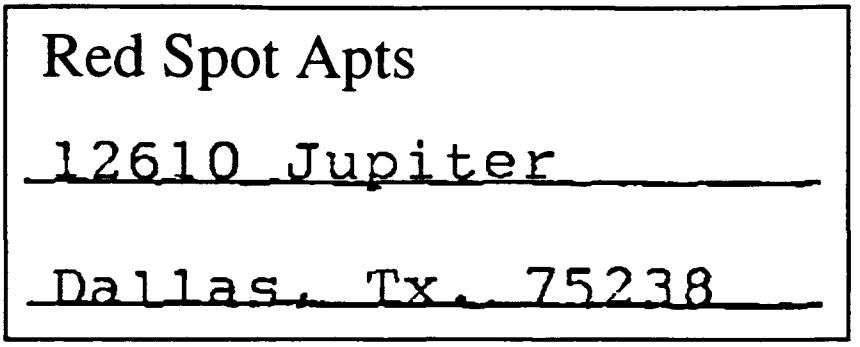



Figure 5.26 - A form address block.

**NR Issue**

Numeral reading issue errors occur in low quality images where the identity of the one or more numerals is questionable. See Figure 5.27 below. In this example, our numeral reader reads the third digit of the primary number as 3. The LOS file claims it is an 8. Since 713 and 718 are legal primary numbers for different range records, database context cannot help. Since we disagree with the LOS, we receive an error for the mailpiece. The proper behavior in this situation is probably to send the mailpiece to 75040.

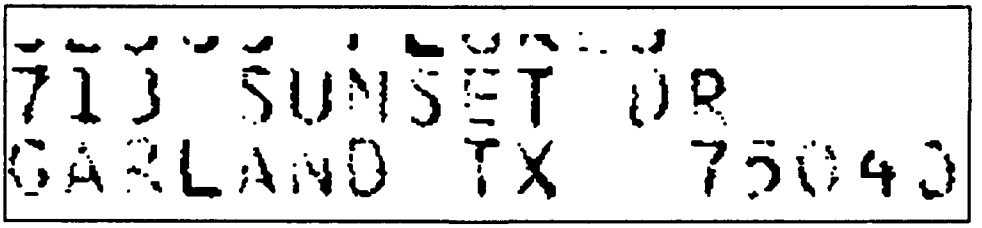



Figure 5.27 - A numeral reading issue error.

**Decision Strategy**

Decision strategy errors occur when the correct ZIP+4 record was among the records proposed by address block verification, but either a different (incorrect) record was chosen or the image was rejected. In the example of Figure 5.28 below, the confidences in the verifications are low and the image is rejected.

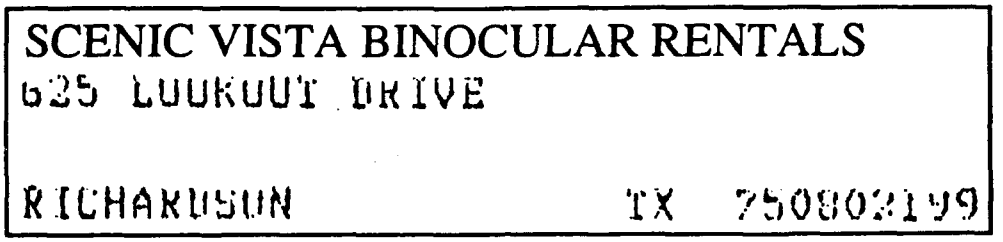



Figure 5.28 - A decision strategy error.

## 5.4. Matching

Figure 5.29 shows the number of times each of the matching error types were observed.

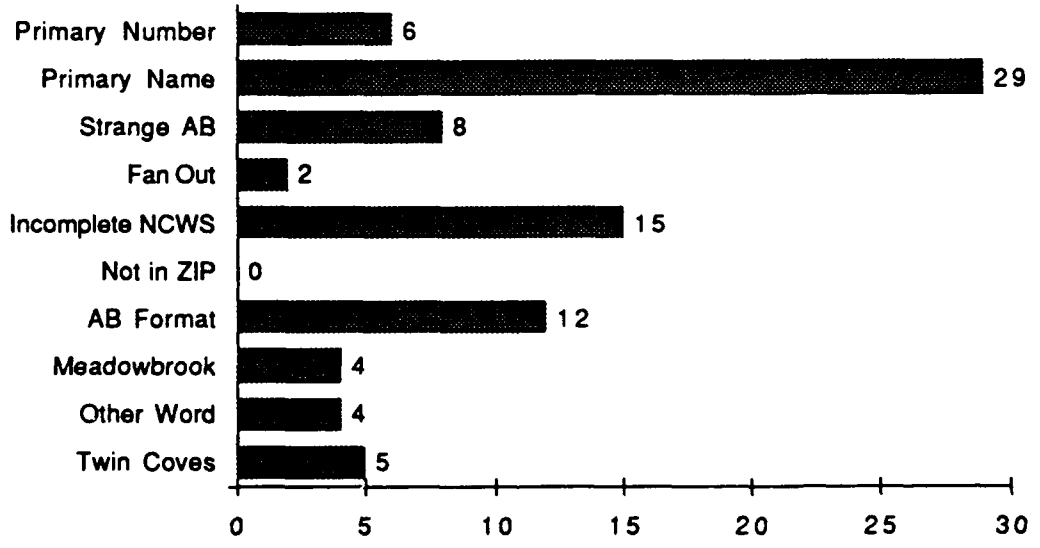



Figure 5.29 - Breakdown of matching errors.

### Primary Number

*Primary number errors* occur when the primary number on the mailpiece does not match the information in the database exactly. Typically an alphabetic character is attached to the end of the primary number when there is no indication of alphabetic characters in the database. As an example, consider Figure 5.30 below. The mailpiece reads "301B." The correct ZIP+4 record has a primary range of 301 - 399 odd. We fail to match because we use the database to tell when to split the field up into numeric and alphabetic parts.

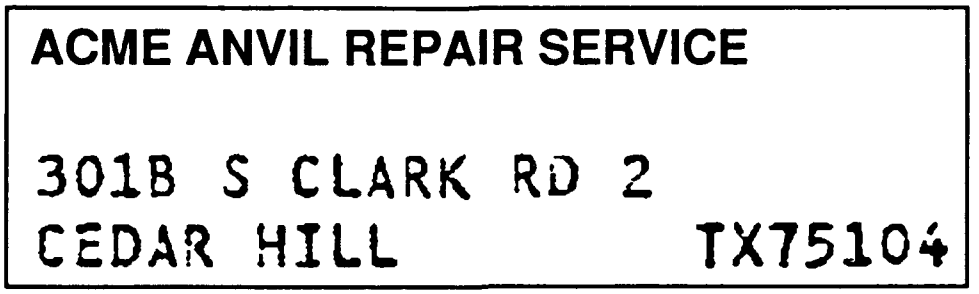



Figure 5.30 - A primary number error.

**Primary Name**

Primary name errors occur when the primary name on the mailpiece does not match the information in the database. Usually this is due to a dropped character, an added character, a spelling error, or an abbreviation. Figure 5.31 shows an example where a letter has been dropped from the primary name. The correct name is "MURRAY."

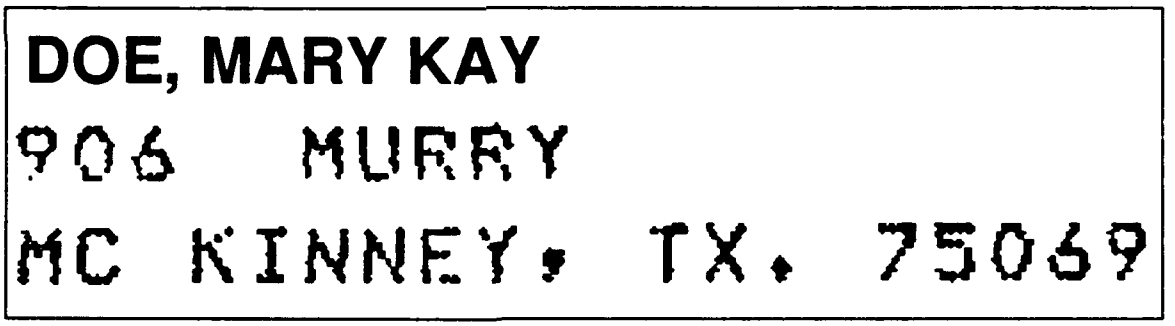



Figure 5.31 - A primary name error.

**Strange Address Blocks**

Strange address block errors occur when the mailpiece contains only secondary name information with no street or PO Box information. See, for example, Figure 5.32 below.

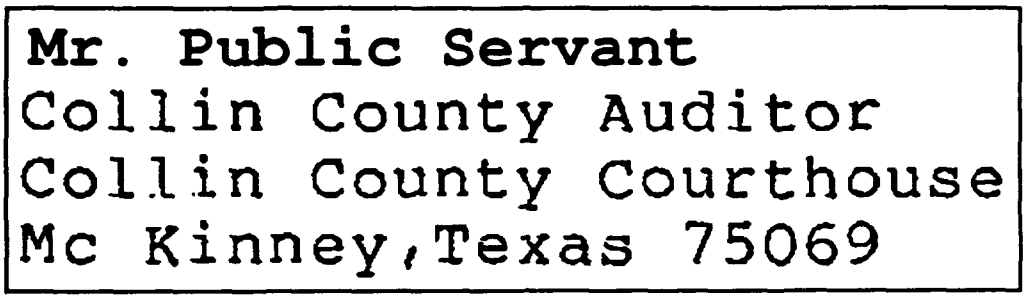



Figure 5.32 - A strange address block error.

**Fan Out**

Fan out errors occur when there are too many ZIP+4 records to be considered and the system reaches its processing time limit. See Figure 5.33 below, an image with primary name "WALNUT." Since WALNUT is a common primary name, the number of target records returned by the query on WALNUT is large. Since the image quality is poor, it is likely that the correct target table record doesn't match any better than the other records. Therefore, the first add-ons which are retrieved might not be from the correct target table record. If the number of other add-ons which are considered first is large, the system will time-out before it retrieves the proper add-on record.

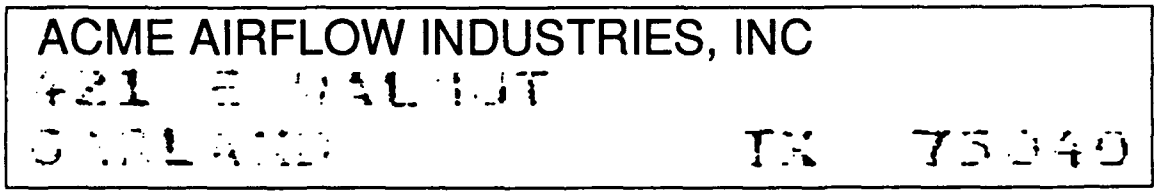



Figure 5.33 - An image with the fan-out problem.

**Incomplete NCWS**

Incomplete NCWS errors occur when the primary number on the mailpiece matches a range in a ZIP+4 record but the walk sequence information for the mailpiece does not contain the specific number. Consider the image in Figure 5.34 below. The primary range for the correct ZIP+4 record has a range of 5600 - 5698 even. Unfortunately, the NCWS says that the only primary numbers in this range which really exist are 5626 and 5620. Because we use the NCWS information whenever it is present, we do not find a match and instead reject the mailpiece.

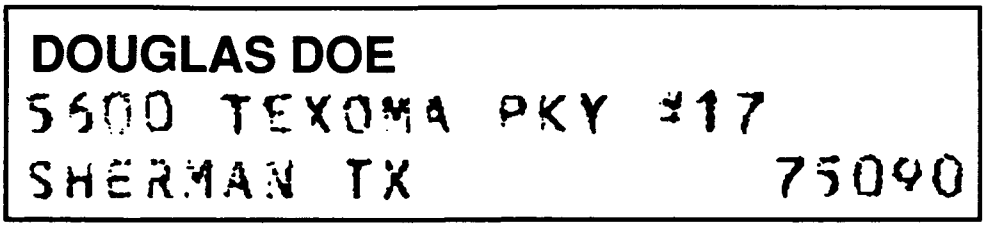



Figure 5.34 - An Incomplete NCWS error.

**Not in ZIP**

Not in ZIP errors occur when the database does not contain any records matching the information on the mailpiece exactly but does contain a close match for the primary name. Figure 5.35 below shows an example. There is no EAST street in Garland, but there is an 814 EASY street in Garland, in ZIP 75042. We respond with this match and receive an error. For the two examples of this which occur, both are LOS-5s in our revised truth. It is probably true that our system is correct for these mailpieces and the LOS and our score should be modified.

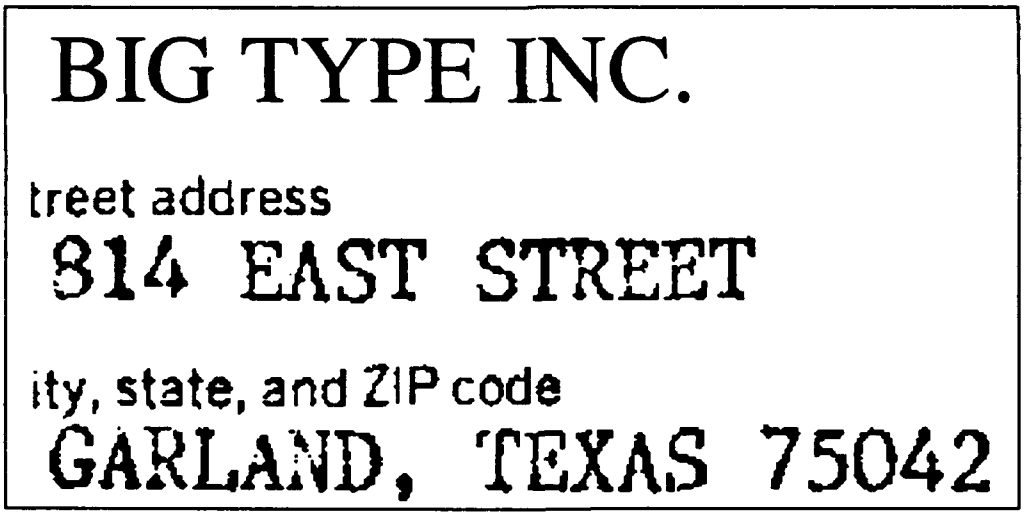



Figure 5.35 - A not in ZIP error.

**AB Format**

Address block format errors occur when the address block structure on the mailpiece is improperly spaced. In contrast to word segmentation errors, these are cases where the patron has omitted spaces between fields. See Figure 5.36 below. In this case, the primary number is directly adjacent to the first primary word.

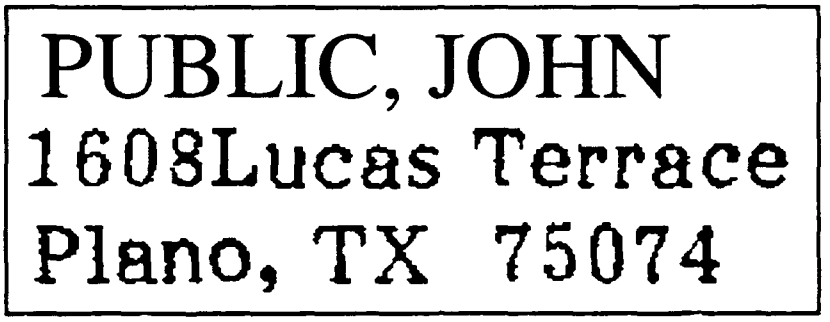



Figure 5.36 - An address block format error.

**Meadowbrook**

Meadowbrook errors occur when the mailpiece contains a street name with two words but the corresponding name in the database is only one word long. See Figure 5.37 as an example. The proper database record for this mailpiece has street name "MAPLECREST." This category can be thought of as a subcategory of primary name errors.

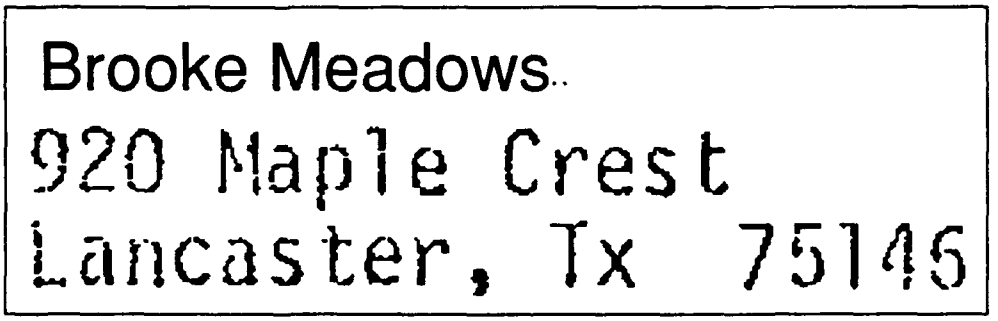



Figure 5.37 - A Meadowbrook image.

**Other Word**

Street recognition system other word errors are related to the method of assigning confidences to potential primary words. For the Phase III test, the confidence was simply the verification score. More realistically, the confidence should take into account prior statistics and the number vs. non probability of the first word on the line of the candidate primary word. An example of this kind of error is shown in Figure 5.38 below. The word "MEDICAL" verifies higher than "FIFTEENTH", so primary records are first retrieved for "MEDICAL." This may lead to timeout if too many records for "MEDICAL" are considered.

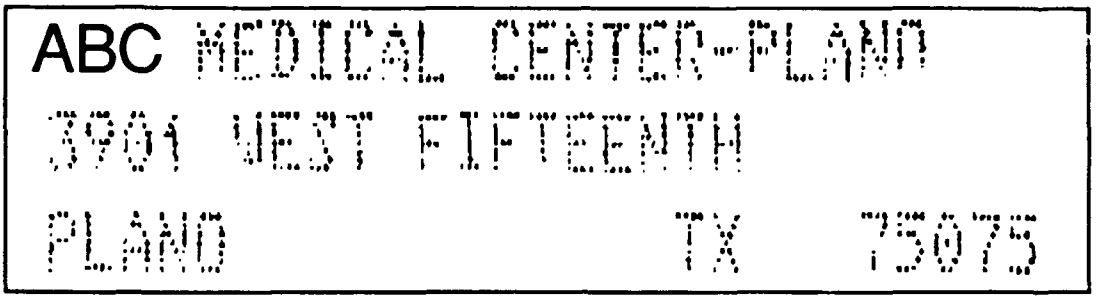



Figure 5.38 - Example of street recognition system other word error.

**Twin Coves**

Twin Coves errors occur when the mailpiece contains a two-word primary, and one or more of the individual primary words is also a primary name by itself. In this situation, due to the way the word verification module assigns confidences, the individual words always verify higher. Consider the image in Figure 5.39. Since "Grove" and "ELM" are both legal primary words by themselves, they will both verify higher than "ELM GROVE." This means that they might receive first consideration in address block verification, and potentially cause timeout before "ELM GROVE" is considered.

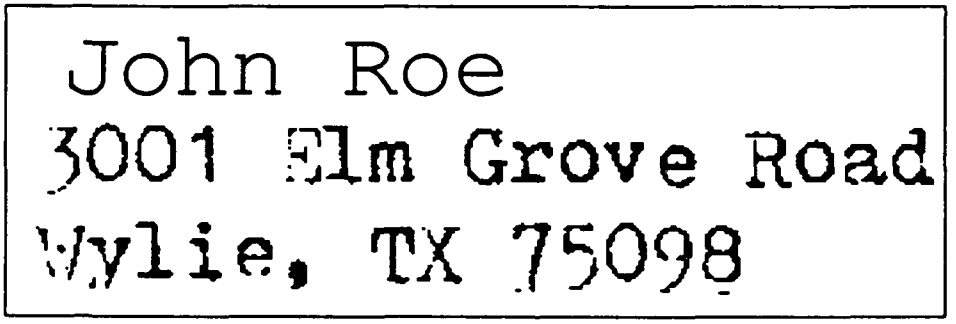



Figure 5.39 - A Twin Coves problem.

## 5.5. Both Image Processing and Matching

This category includes mailpieces which suffered from both image processing and matching problems. Also included are two other categories: bad parses and other. Figure 5.40 shows the number of times each of these types of errors were observed.

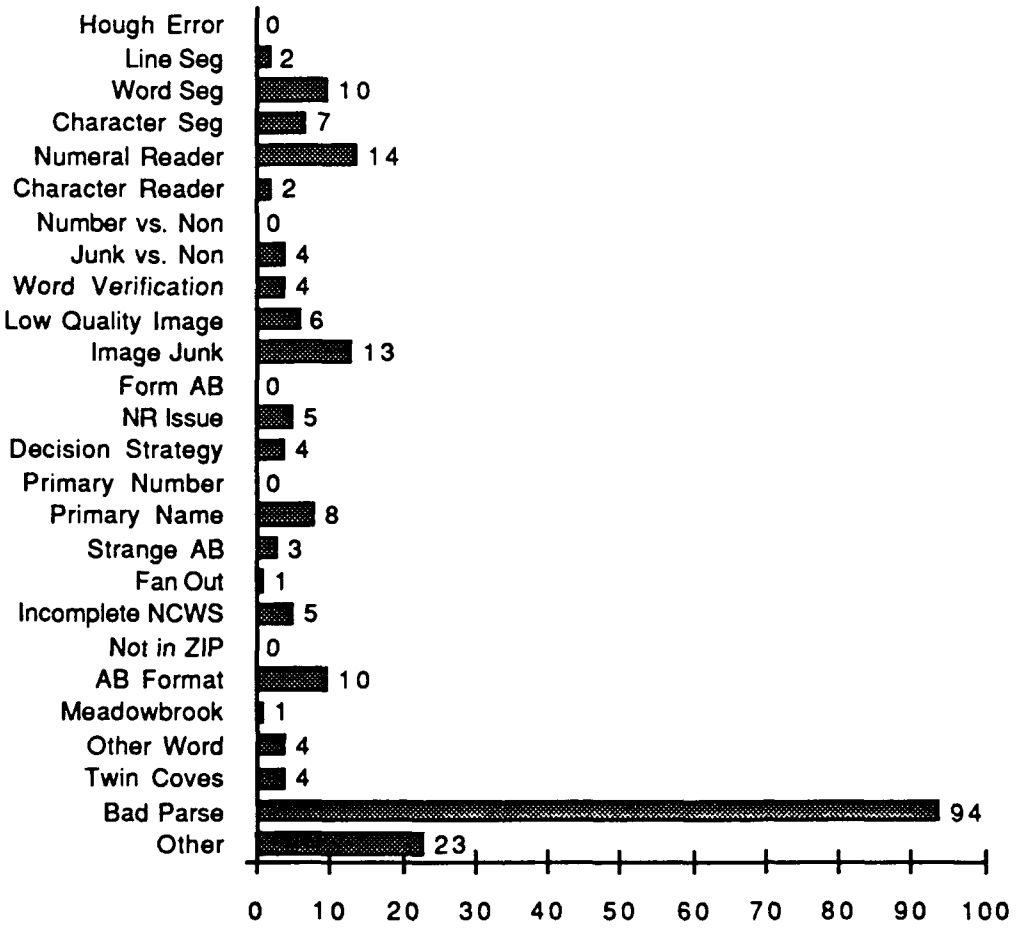



Figure 5.40 - Breakdown of both image processing and matching, bad parse, and other errors.

### Bad Parse

Bad Parse errors occur when the mailpiece image is assigned the wrong mailpiece type. Figure 5.41 below shows a PO Box image which has been incorrectly parsed as a street.

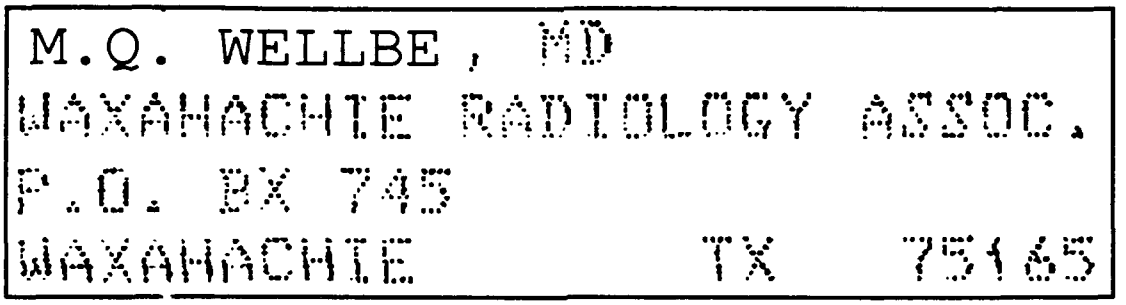



Figure 5.41 - A PO Box image parsed as a street.

## 5.6. Error Summary

The combination of Image Processing, Matching, and Both Image Processing and Matching Errors results in an itemization of all of the system errors as shown in Figure 5.42. This Figure summarizes Figures 5.14, 5.29, and 5.40.

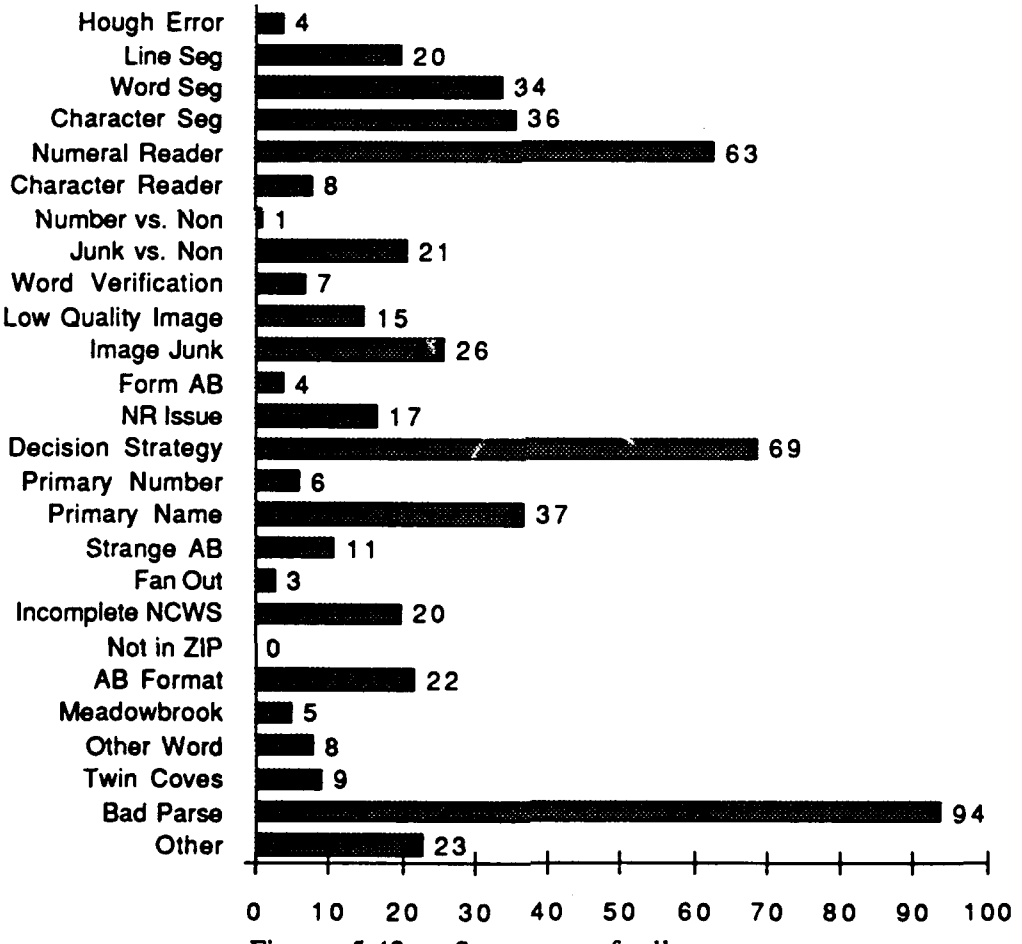



Figure 5.42 - Summary of all errors.

## 5.7. Prognosis

In this section we present our prognosis for the errors discussed in the previous sections. Errors have been divided into three groups: easy, middle, and hard. The easy group contains errors that can be fixed with minimal effort. The middle group contains errors that can be fixed with a moderate effort. The hard group contains errors for which a solution is not immediately obvious and which will require significant effort to solve. Table 5.1 shows the breakdown of the errors into the four major categories and the prognosis for the errors in those categories. Table 5.2 shows the breakdown of the errors based on the encode level and the prognosis for the errors based on the encode level.

Table 5.1 - Error Type Prognosis

| | Easy | Middle | Hard | Total |
|---|---|---|---|---|
| LOS-5 | 0 | 0 | 143 | 143 |
| Image Processing | 31 | 148 | 19 | 198 |
| Matching | 2 | 45 | 29 | 76 |
| Both IP & M | 22 | 104 | 1 | 137 |
| Total | 55 | 297 | 202 | 554 |

Table 5.2 - Encode Level Prognosis

| | Correct | Easy | Middle | Hard | LOS-5 | Total |
|---|---|---|---|---|---|---|
| 5 Digit | 207 | 15 | 108 | 27 | 94 | 451 |
| 9 Digit | 128 | 4 | 32 | 3 | 1 | 168 |
| Reject | 122 | 38 | 157 | 29 | 48 | 394 |
| Total | 457 | 57 | 297 | 59 | 143 | 1013 |

## 6. SYSTEM TIMING ANALYSIS

A complete timing test was performed on the context system at the end of Phase III. Each major module in the system was analyzed along with the major support modules of the system. This data was generated using 398 images from the Phase III test set. The pie chart in Figure 6.1 shows the break down for the major modules of the system.

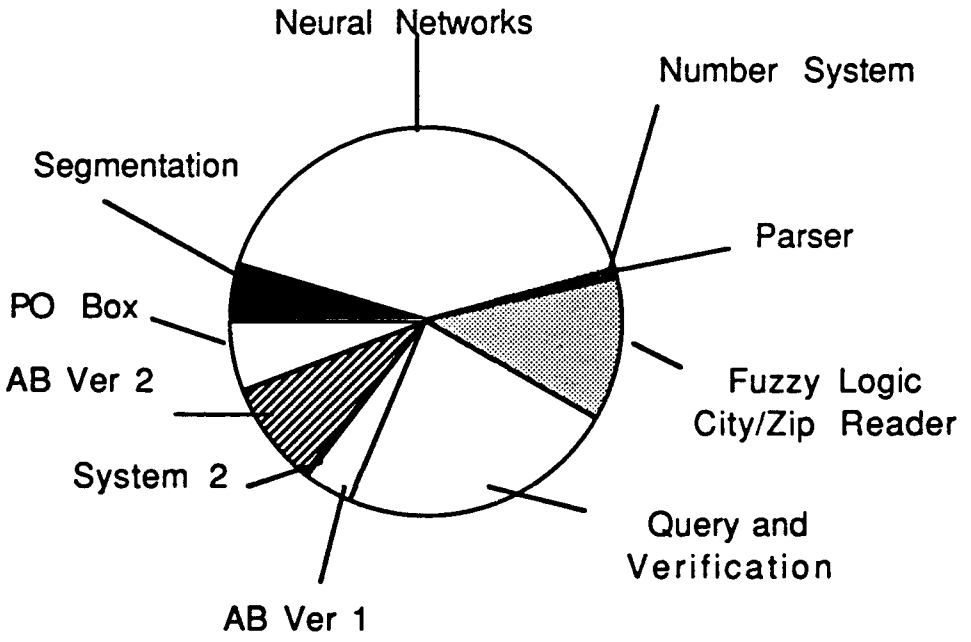



Figure 6.1. - Timing Study Overview

The chart shows that segmentation combined with the neural networks account for approximately half of the time. These modules combined make up the preprocessor to the system. These two modules are a one time process with their benefits being used multiple times throughout the system. The performance of the segmentation modules is shown in Table 6.1. The average number of outputs for the word and character segmentation modules is higher than the average number of words and characters in an address block image. This results from both modules outputting multiple segmentations. This increases the time in the neural networks and other processes downstream, but adds significantly to system performance.

Table 6.1. - Segmentation System Analysis

| | Average Time (seconds) | Average Number Output |
|---|---|---|
| Line | 13.12 | 3.4 |
| Word | 1.9 | 15.6 |
| Char | 29.8 | 403.3 |
| Total | 44.8 | |

The time shown in Figure 6.1 for the neural networks may be inflated due to the fact that the networks were run on each of the characters in the image during the test while in actual processing of an image only the networks that are needed for a particular character image are run. Table 6.2 shows the average time of all the networks along with the times for the individual networks. This table shows that the two most time consuming networks are the channel network and the feature extraction.

Table 6.2. - Neural Network Time

| | Average Time (seconds) |
|---|---|
| Feature Extract | 41.8 |
| Channel Net | 182.6 |
| Numeral Net | 37.8 |
| Numeral & "#" | 39.5 |
| Number vs. Non | 36.3 |
| Junk vs. Non | 35.8 |
| Total | 373.8 |

The two most time consuming modules in the processing part of the system are the fuzzy logic ZIP/city reader and the query and verify system with their times shown in Table 6.3. The main reason for the large amount of time spent in the ZIP/city reader is its large number of calls to word verification. On the average, the ZIP/city reader makes 31,847 calls to word verification out of a total of 35,473 calls during the processing of an average image. The word verification module can process 500 images per second, or 30,000 per minute after all of the neural networks have been run. This means that the bulk of the time spent in the ZIP/city reader is used to do word verification. The current implementation of the ZIP/city reader is exhaustive in its verification of cities and ZIPs. By using lexicon pruning and information about verification confidences and thresholds, we feel that we can drastically reduce the number of calls made to word verification. The large amount of time spent in the query and verification module results from the structure that the module uses. The structure is very large and must be continually maintained.

Table 6.3. - Query and Verify and City/ZIP Reader Times

| | Average Time (seconds) |
|---|---|
| Query and Verify | 210.6 |
| Fuzzy Logic ZIP/City Reader | 98.8 |

The remainder of the modules in the system are not considered major time consuming modules. Their times are shown in Table 6.4.

Table 6.4. - Other Module Times

| | Average Time (seconds) |
|---|---|
| Parser | 4.7 |
| Number System | 2.9 |
| AB Verification I | 36.14 |
| System II | 2.5 |
| AB verification II | 79.6 |
| PO Box Processing | 18.53 |

Figure 6.2 shows a overview of the system modules (identical to Figure 2.1) with the average processing times attached to each module. The average total time for a street image is 864 seconds and the average total time for a PO Box image is 587 seconds. All modules in the system are being thoroughly analyzed in an attempt to reduce processing time. There are some apparent changes like reducing the number of calls that the ZIP/City reader makes to word verification, but other improvements are not as quick and easy. We anticipate that some changes can be made to improve system throughput, but additional capabilities added to the system in Phase IV will add additional processing time to the overall system.

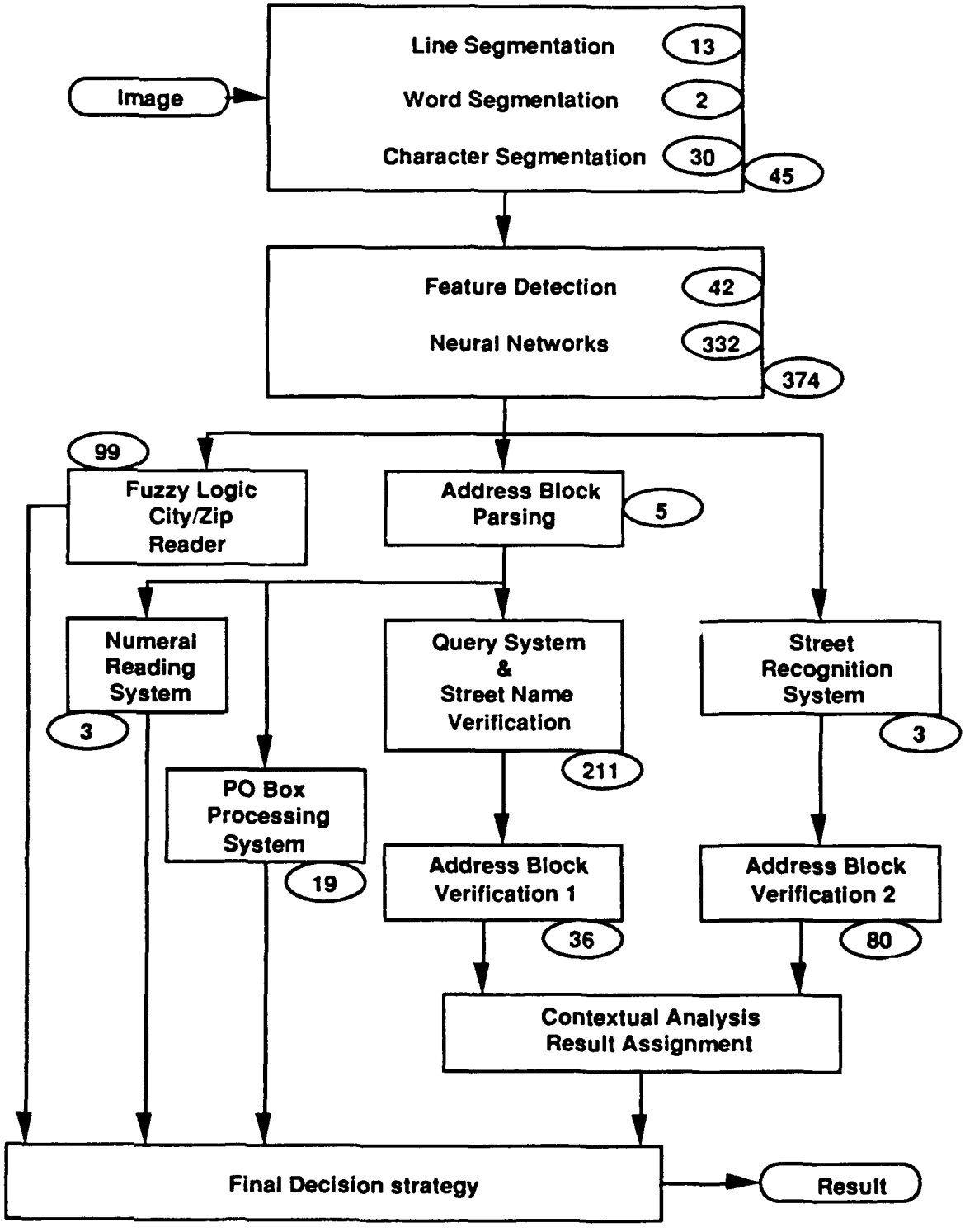



Figure 6.2 - Timing of system modules (all times in seconds).

Also of interest are some statistics regarding the usage and output of various modules. Table 6.5 shows the average, maximum, and minimum number of lines, words, and characters output by the segmentation process for a mailpiece.

Table 6.5 - Segmentation Statistics

| | Average | Maximum | Minimum |
|---|---|---|---|
| Number of lines | 3.4 | 6 | 2 |
| Number of words | 15.6 | 38 | 2 |
| Number of characters | 430.3 | 1924 | 93 |

Table 6.6 shows the average and maximum number of word verifications carried out by various modules for a mailpiece.

Table 6.6 - Word Verification Usage (number of calls)

| | Average | Maximum |
|---|---|---|
| Parser | 1162 | 3,320 |
| Numeral Read | 140 | 252 |
| Fuzzy Logic Read | 31,847 | 112,966 |
| Query&Verify | 2626 | 5,114 |
| AB Verify #1 | 129 | 632 |
| Street Recognition | 115 | 878 |
| AB Verify #2 | 529 | 3,750 |
| PO Box | 152 | 252 |
| Total | 35,474 | 116,212 |

Word Verification Time Usage:

| | |
|---|---|
| Average time/AB-image: | 69 secs |
| Max time/AB-image: | 267 secs |
| Average time/call: | 1.9 msecs |

Table 6.7 shows the average and maximum number of database accesses for a mailpiece.

Table 6.7 - Database Access Statistics

| | Average | Maximum |
|---|---|---|
| Street Records Accessed | 17 | 108 |
| Range Records Accessed | 180 | 2833 |

# 7. PATRON ERROR ANALYSIS

An analysis of patron errors on images from the Phase III test was conducted. The purpose of this analysis was to characterize the problems due to patron errors in the input to the contextual analysis system, and to evaluate how well it handles erroneous information.

## 7.1. ZIP Code

The ZIP code had more patron errors than any other piece of information on the mail piece. The Table 7.1 shows the Phase III system response to images with ZIP code errors broken down by result and by which module in the system produced the result. The first column shows that there were a total of 93 errors detected in the ZIP code. These errors occurred in both the five digit portion and the four digit add-on. Of the 93 errors the system processed the mail piece 51 times with a correct four digit add-on. All of these recoveries were made in the contextual analysis system (see column labeled SYS). The partial parser (also called the fuzzy logic ZIP code reader) was hurt the most by the errors, with 11 errors in the five digit code. This is caused by the city ZIP comparisons on which the partial parser depends.

Table 7.1 - Patron Errors in ZIP Code

| Result | PP | SYS | CITY NN3 | PP5 DD | NN9 | 5DD | NN5 | None | Total |
|---|---|---|---|---|---|---|---|---|---|
| 9-digit Direct | 0 | 22 | 0 | 0 | 0 | 0 | 0 | 0 | 22 |
| 9-digit HA | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| 9-digit S | 0 | 21 | 0 | 0 | 0 | 0 | 0 | 0 | 21 |
| 5-digits | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 3-digits | 0 | 0 | 2 | 0 | 0 | 0 | 3 | 0 | 5 |
| Error Add-on | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| Error 5-digit | 11 | 2 | 0 | 0 | 0 | 0 | 3 | 0 | 16 |
| Unresolved | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 | 13 |
| Total | 15 | 57 | 2 | 0 | 0 | 0 | 6 | 13 | 93 |

## 7.2. Street Name

The street name on the mail pieces also had a significant number of patron errors. Table 7.2 shows that there was a total of 48 errors detected. Of these 48 errors, 11 were correctly processed with a four digit add-on. As with the ZIP code errors, the recovery from errors came from the contextual analysis system. Most of the remaining errors were processed at the five digit ZIP code level.

Table 7.2 - Patron Errors in Street Name

| Result | PP | SYS | CITY NN3 | PP5 DD | NN9 | 5DD | NN5 | None | Total |
|---|---|---|---|---|---|---|---|---|---|
| 9-digit Direct | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9-digit HA | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 9-digit S | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 5-digits | 23 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 27 |
| 3-digits | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 |
| Error Add-on | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| Error 5-digit | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 3 |
| Unresolved | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| Total | 25 | 14 | 2 | 0 | 0 | 0 | 5 | 2 | 48 |

Table 7.3 lists all of the 48 errors detected. These errors are divided into four categories: meadowbrook, misspellings, abbreviations and wrong. The meadowbrook problem is one in which the database contains a one word primary while the patron prefers to write it as two words. The abbreviation problem occurs from both the patron and the database being abbreviated.

Table 7.3 - Street Name Patron Errors

| Database | Mail Piece |
|---|---|
| *Meadowbrook Problem* | |
| BLUFFVIEW | BLUFF VIEW |
| MEADOWCREEK | MEADOW CREEK |
| PEPPERIDGE | PEPPER RIDGE |
| LAKEHILL | LAKE HILL |
| MAPLECREST | MAPLE CREST |
| BRIARCOVE | BRIAR COVE |
| MCDANIEL | MC DANIEL |
| PINETREE | PINE TREE |
| HILLVIEW | HILL VIEW |
| MOSSVINE | MOSE VINE |
| *Misspelling* | |
| ASBURY | ASOURY |
| BUCKNER | BUCHNER |
| MURRAY | MURRY |
| WILLIAMS | WILLIAM |
| DAVIS | DAVIES |
| NICHOLS | NICHOLAS |
| FISHER | FISCHER |
| KIRBY | KERBY |
| SHORE | WESTSHORE |
| GARDENGATE | GARDENGAGE |
| WAINWRIGHT | WAIHWWRIGHT |
| RAINIER | RANIER |
| AZZURA | AZZURRA |
| BOBBY | BOBBIE |
| WINDWARD | WINWARD |
| DRUMCLIFFE | DRUMCLIFF |
| CHENNAULT | CHENALT |
| CRESCENT | CRESENT |
| HILLCREST | MHILLCREST |
| HARBINGER | MARSINGER |
| SENECA | SENICA |
| FOREST HILL | FORREST HILL |
| SOUTHWYND | SOUTHWIND |
| LAS COLINAS | LOS COLINAS |
| BAHAMA | BAHAM |
| *Abbreviation* | |
| BIG TOWN SC | BIG TOWN MALL |
| EDWARDS CHURCH | EDWARDS |
| NORTHWEST | NW |
| INTERSTATE 45 or I 45 | INT 45 |
| INTERNATIONAL | INTL |
| DRIFTWOOD VILLAGE SC | DRIFTWOOD VILLAGE SHOPPING CENTER |
| PARK SQUARE | PARK SQ |
| INTERSTATE 35 or I 35 | IH-35E |
| GREAT SOUTHWEST | GSW |
| *Wrong* | |
| HIGHWAY 80 | HIGHWAY 30 |
| I 35 SERVICE | SERVICE |
| INTERSTATE 35 SERVICE | SERVICE HWY 35 |
| TEXOMA | HIGHWAY 75 |

## 7.3. City Name

The city name was also studied for patron errors. Table 7.4 gives the detailed breakdown. The city name had only 17 patron errors. Of the 17 errors, 7 were correctly processed with a nine digit ZIP code. The contextual analysis system was once again the module that recovered. Table 7.5 contains a list of the city name errors which occurred. The patron either simply spells the city name incorrectly or puts a totally incorrect city name on the mail piece.

Table 7.4 - Patron Errors in City Name

| Result | PP | SYS | CITY NN3 | PP5 DD | NN9 | 5DD | NN5 | None | Total |
|---|---|---|---|---|---|---|---|---|---|
| 9-digit Direct | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9-digit HA | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 9-digit S | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| 5-digits | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 3 |
| 3-digits | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| Error Add-on | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Error 5-digit | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| Unresolved | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 |
| Total | 3 | 8 | 1 | 0 | 0 | 0 | 2 | 3 | 17 |

Table 7.5 - City Name Patron Errors

| Database | Mail Piece |
|---|---|
| UNIVERSITY PARK | UNIV PARK |
| RED OAK | ORVILLE |
| DENISON | DENNISON |
| THE COLONY | COLONY |
| PLANO | PLANTO |
| SHERMAN | SHERMNA |
| DALLAS | ADDISON |
| HUTCHINS | HUTCHINGS |
| RED OAK | GLENN HEIGHTS |
| DALLAS | ADDISON |
| CARROLLTON | CAROLLTON |
| DALLAS | ADDISON |
| GRAND PRAIRIE | BRAND PRAIRIE |
| ROYSE CITY | ROCKWALL |
| DALLAS | VICKERY |
| DALLAS | ADDISON |

## 7.4. Secondary Name

In many cases, the secondary name can be the deciding piece of information in the processing of a mail piece. The analysis of secondary name errors focused on the mismatches between the mail piece and the database. This does

not necessarily mean that the patron is incorrect. Many times the information in the database is incomplete or abbreviated so as to fit into the field of 25 characters. Figure 7.1 shows some examples of the database abbreviation problem. It can be very difficult to match the mail piece to the database as can be seen in the confusing abbreviation in the last address in Figure 7.1.

| | |
|---|---|
| Database Record: | |
| | DRI<br>9713 HARRY HINES BLVD<br>DALLAS, TX 75220 |
| Mail Piece: | |
| | DALLAS REHABILITATION<br>9713 HARRY HINES BLVD<br>DALLAS TX 75220 |
| Database Record: | |
| | CORSICANA H S<br>3900 W HWY 22<br>CORSICANA, TX 75110 |
| Mail Piece: | |
| | CORSICANA HIGH SCHOOL<br>WEST HIGHWAY 22<br>CORSICANA, TX 75110 |
| Database Record: | |
| | CIFSC<br>600 E LAS COLINAS BLVD 1400<br>IRVING, TX 75039 |
| Mail Piece: | |
| | CONNECTICUT GENERAL LIFE INS. CO.<br>CIGNA TOWER, #1400<br>600 EAST LOS COLINAS BLVD.<br>IRVING, TX 75039 |

Figure 7.1 - Database Abbreviation Problem Examples

Many times there are words that match between the database and the mail piece. But, some words carry little information, such as "Corp." or "Inc.," while others are very important. Figure 7.2 shows some examples of matches using "important" words. The second address block in Figure 7.2 has a strong word match of "GRAPHICS" between the database and the mail piece, but it is not clear if the database and the mail piece should match.

| | |
|---|---|
| Database Record: | |
| | KAFM, KZPS, FAAM<br>15851 DALLAS PKY 1200<br>DALLAS, TX 75248 |
| Mail Piece: | |
| | VICKI ROBBIN KZPS<br>15851 DALLAS PKWY. #1200<br>DALLAS, TX 75248 |
| Database Record: | |
| | WHITE GRAPHICS<br>925 AVENUE N<br>GRAND PRAIRIE, TX 75050 |
| Mail Piece: | |
| | AVERY GRAPHIC SYSTEMS GRAPHICS<br>925 AVE N<br>BRAND PRAIRIE, TX 75050 |

Figure 7.2 - Important Word Problem Examples

The final type of secondary name problem is simply where that patron types something unexpected. Figure 7.3 shows an example of an unexpected patron error. The database and the mail piece clearly match, but how can the "SMTIDWELL" be explained. It is unclear if SM is an abbreviation for Samuel, or if Samuel's middle initial is M. Matching this type of patron error is a difficult problem.

| | |
|---|---|
| Database Record: | |
| | SAMUEL TIDWELL<br>12222 MERIT DR 1450<br>DALLAS, TX 75251 |
| Mail Piece: | |
| | SMTIDWELL & ASSOCIATES<br>D L KARAPETIAH<br>12222 MERIT DR SUITE #1450<br>DALLAS, TX 75251 |

Figure 7.3 -Patron Error Example

The matching of secondary information is a very difficult task. A loose matcher must be constructed so all of the above errors can be included, but not too loose so as to include things like our "GRAPHICS" example that may not really match.

## 8. PHASE IVa PLAN

Figure 8.1 shows the planned Phase IVa contextual analysis system. Figure 8.2 shows the schedule of tasks for Phase IVa. Figure 8.3 shows the linkage between the tasks and the errors discussed in Section 5.

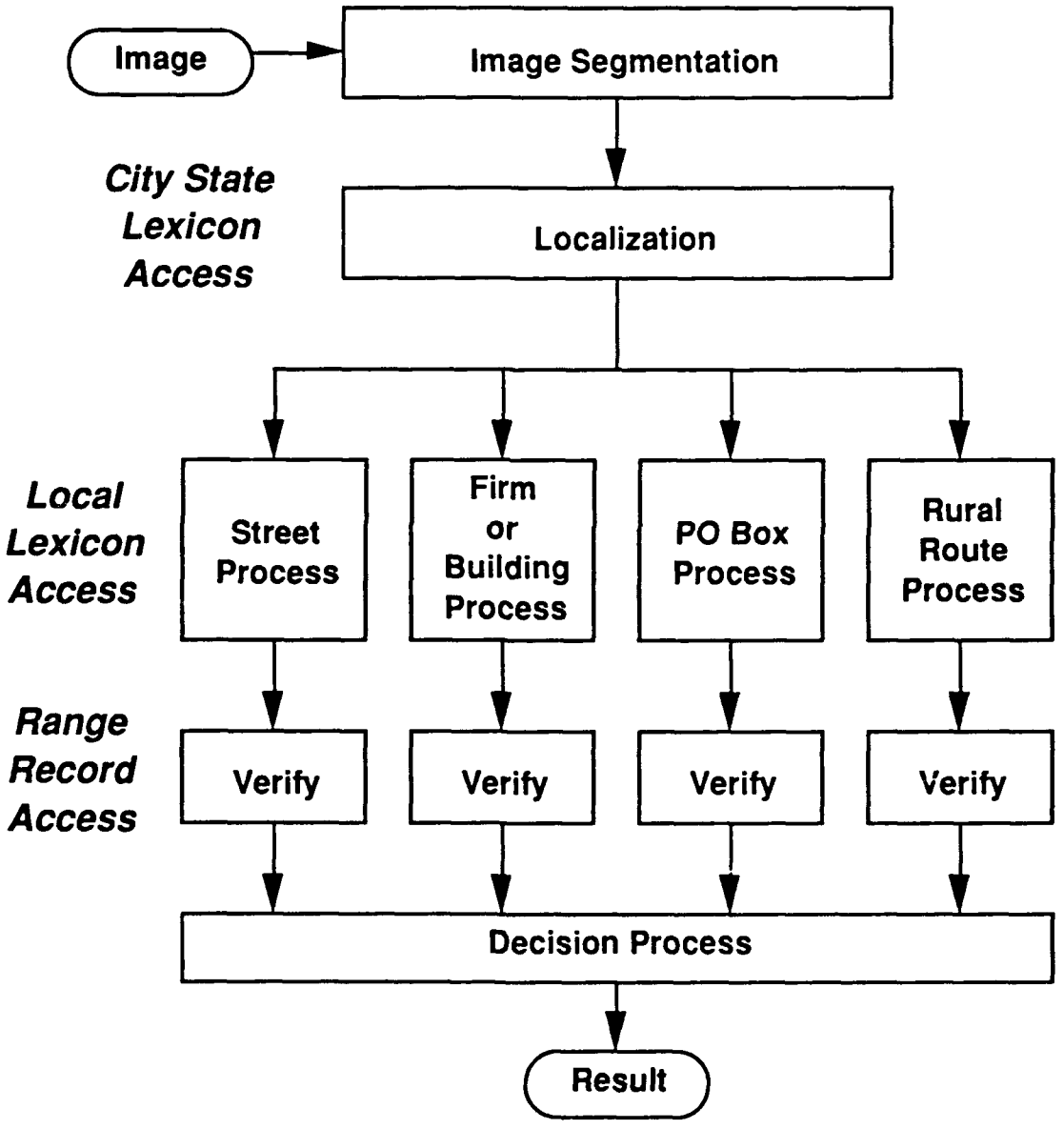



Figure 8.1 - New contextual analysis system

Figure 8.2 - Phase IV task schedule

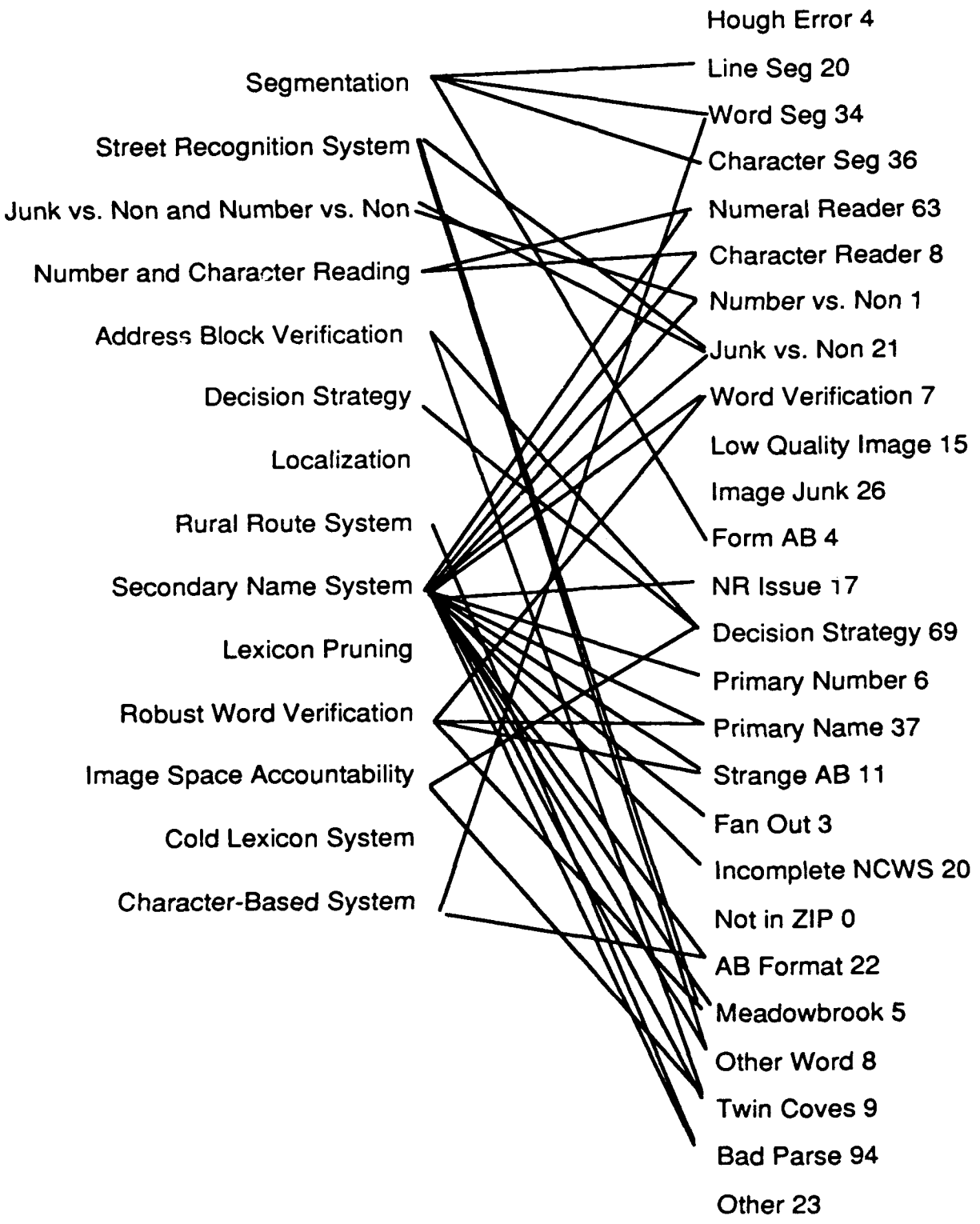



Figure 8.3 - Matching of tasks with errors

The tasks for Phase IVa can be divided into 5 groups: Phase IV system, refinements, new modules, new techniques, and supplemental tasks. These groups of tasks are discussed in the following sections.

## 8.1. Phase IV System

This task involves the construction of the 25-SCF database, the reconstruction of the end-to-end system, and the testing of the system.

## 8.2. Refinements to Existing Modules

Based on the error analysis, several system modules will be enhanced.

### Segmentation

The line segmentation algorithm will be improved so that broken or faint lines are segmented properly. Truth information will be used to segment words into characters in order to develop a dataset of cleanly segmented characters. A new character normalization technique should improve character recognition results from the neural nets. A new approach for character segmentation of address blocks with fixed-width characters will be used.

### Street Recognition System

The street recognition system will be improved.

### Neural Network Refinement

This task involves the gathering of a new dataset of cleanly segmented characters through the use of truth-based segmentation and with a new approach to character normalization. Some new features will be evaluated and then the neural nets will be re-trained.

### Address Block Verification

The first enhancement to the address block verification module will be the unified treatment of image space accountability and database record accountability. This involves accounting for everything found on the mailpiece and everything stored in a database record when verifying an address. The second enhancement is the incorporation of secondary name verification. The third enhancement is to ensure that verifications for street, firm/building, PO Box, and rural route mailpiece are compatible so that the decision strategy can make correct decision between alternative hypotheses.

### Decision Strategy

The decision strategy will be enhanced to combine the best aspects of the Gaussian-based approach and the Dempster-Shafer approach into a single strategy. Also, it will be necessary to handle firm/building, PO Box, and rural route verifications as well as street verifications.

## 8.3. New Modules

Three new modules will be added to the contextual analysis system.

### Localization

The localization module is necessary because the system needs to handle 25 SCFs and eventually must handle the entire country. Localization is used to limit the size of the area under consideration for more detailed queries. The approach is to parse for the city, state, and ZIP and to find hypotheses for each word individually. Then, a reconciliation process produces a list of ranked hypotheses for city/state/ZIP triplets. An alternate localization process will examine the street name and provide city/state/ZIP hypotheses if the street name is one of the many street names with a low frequency of occurrence. This approach is referred to as the cold lexicon approach because the street names in the lexicon are rare but very useful for localization.

### Rural Route System

The rural route system will handle rural route mailpieces which were not handled in the Phase III system. The approach is similar to that used for PO Boxes except that it must deal with route numbers and route box numbers.

### Secondary Name System

The secondary name system will use secondary names when the information is available as a supplement to other address information and will also be useful when there is no other address information on the mailpiece. The primary issue of concern is inexact matching of secondary names.

## 8.4. New Techniques

### Lexicon Pruning

New lexicon pruning techniques must be developed because lexicon sizes increase as we move toward the national database. It will also be useful for the cold lexicon approach to localization and for local lexicons for street names. A number of approaches are being considered: word length, character recognition results, match of N of M characters, allowing dropped or added characters, n-grams, character equivalence classes, and pre-computed structures.

### Robust Word Verification

The goal of robust word verification is inexact word matching and the benefits of robustness will be seen throughout the system. Some approaches being considered include: sequence matching with dropped and added characters, rules for abbreviation, and n-gram signatures.

### Image Space Accountability

Image space accountability is required because there are cases where information on the mailpiece has been ignored but if used it would lead to a different decision. This is often the case for street names with multiple words where one of the words is a valid street name by itself.

### Cold Lexicon System

The cold lexicon system will be used for localization. It will use character recognition information to search for the street name in a large lexicon of rare street names.

### Character-Based System

The character-based system will bypass problems with word segmentation by looking at entire lines and segmenting words as they are recognized.

## 8.5. Supplemental Tasks

### Text Matching Approaches

When matching information from a mailpiece to the postal directory, various forms of inexact matching must be employed. Inexact matches can come from misspelled or abbreviated words, rearranged, missing, or spurious words, or from the use of acronyms and synonyms. This problem is of major significance in the development of algorithms which make use of secondary names. In this task we propose to study existing approaches to the text matching problem, and develop methods for use in the address matching module. Topics for study include the Viterbi algorithm which makes use of letter transition probabilities the Soundex method for "sound-alike" matching, the use of n-gram signatures for matching secondary names and rule-based approaches for abbreviations and acronyms.

### National Database Analysis

The ERIM contextual system makes use of a number of different queries or access methods into the database. Still more queries can be defined. Also, different combinations of queries may be used. Each query (access method) has a certain cost in storage, and a certain benefit in number of mailpieces which depend on that access route to be encoded. In this task we propose to make a cost/benefit tradeoff study of the national database, with respect to these issues. This study must find ways of weighing the costs and benefits of each structure in the database. We anticipate that the benefits of some queries will be high in some ZIP codes and low in others, leading to a database structure which is dependent on the information in each zone. Also, the study should take into account the rates of detection available from the image processing side of the system, and the usage frequencies found in the mail stream. The interaction of various database compression approaches with query structure will also be investigated.

### ASCII Address Matching

In this task we will construct a contextual analysis system which uses the truth information from the address blocks as input. Initially, the system will be based on the Phase III system. The goal is to investigate contextual analysis problems without the interference of image processing problems. The performance of the system will be the best-case benchmark for the Phase IVa contextual analysis system. The failures of the system will be analyzed and new strategies will be developed based on the analysis. The primary topics of investigation are patron errors, database errors, database query utility, and the impact of zip translation information.

### Near Real Time System

It is clear that the goal of real time contextual analysis is becoming increasingly important. In this task we propose to study the existing contextual algorithms, and explore hardware systems for implementing these algorithms in near real time (10 mailpieces per minute.) The goal of this task is to define a hardware system for this purpose. Two major benefits of this task are expected. The first is to inform the algorithm development team about time expenditure issues, and acquaint them with the process of moving algorithms to a hardware platform. The second is the benefit of a system which can process images at a near real time rate, thereby making more extensive algorithm testing possible. Adequate testing and error analysis is necessary to drive the algorithm development process toward a successful implementation for the ARU prototype.

### Analog VLSI Support

The possible availability of analog VLSI technology for use in the ARU necessitates special investigations of approaches to address interpretation which could make use of this technology. Of particular interest is the notion of a "sliding window" neural network which could be used to read text without the need for character segmentation. We envision an approach which takes isolated, height-normalized lines as input. A recognizer would be trained to recognize any character centered in a window sliding across this line image. At each position, a feature detection and classification step would occur. If a character is centered under the window, its identity would be expected to be output. If the window is centered over a gap between letters, then a reject signal would be expected. This task would experiment with the design and training of such a recognition system. One focus of the task is the development of a feature set which is robust in the presence of interference from characters next to the centered character.

## 9. CONCLUSION

The results of this phase of research demonstrate that significant benefits can be obtained through the use of contextual analysis. A thorough analysis of system errors has clarified many of the issues that need to be addressed in order to improve the digit encode rate. A plan for the next phase which addresses these issues has been presented.