Res. Rep. Fac. Eng. Mie Univ., Vol. 21, pp. 37−44 (1996)

Original Paper

Handwritten Address Interpretation Using Extended Lexicon Word Matching

Fumitaka KIMURA, Yasuji MIYAKE and Malayappan SHRIDHAR*
(Department of Information Engineering)

(Received September 13, 1996)

Abstract

In this paper, a postal address interpretation system using an extended lexicon word matching is described. The extended lexicon word matching is an extension of a lexicon directed word matching algorithm [1],[2], which allows the lexicon to include lexicon words with variables (wild cards). The extended lexicon word matching is utilized for all word recognition tasks including ZIP code recognition, street number recognition, street name recognition, and POBox number recognition. In a performance evaluation test with 1600 handwritten address images, the error rate was 1.12% with 50.19% encode rate.

Key words : word recognition, document analysis, ZIP code recognition, postal address recognition

1. Introduction

The handwritten address interpretation system consists of subsystems for preprocessing, ZIP code line recognition, street line recognition, POBox line recognition, and determination of delivery point code (DPC) (Fig.1).

The preprocessing subsystem applies tilt correction, line segmentation, slant correction, and word pre segmentation to address block images. The ZIP code line recognition subsystem generates several ranked ZIP code candidates. The subsystems for the preprocessing and the ZIP code line recognition are described in [3].

The street line recognition subsystem generates several ranked pairs of street number and street name for given 5-digit ZIP code. If the top candidate pair is accepted with sufficient confidence it is sent to the DPC determination subsystem together with the 5-digit ZIP code.

The POBox line recognition subsystem generates several ranked POBox number for given 5-digit ZIP code. If the top candidate is accepted with sufficient confidence it is sent to the DPC determination subsystem with the 5-digit ZIP code.

* University of Michigan-Dearborn, USA

If the top candidate is 5-digit unique ZIP code with sufficient confidence, it is encoded directly in the DPC determination subsystem. If the top candidate is 9-digit on mail piece with sufficient confidence, it is also directly encoded to DPC in the DPC determination subsystem.

The DPC determination subsystem encodes given information from each subsystem to a DPC. If no valid DPC is obtained, and the 5-digit ZIP code has sufficiently high confidence, it is accepted. Otherwise it is rejected.

2. Street Line Recognition Subsystem

2.1 Street Number Location and Recognition

Street number is recognized by an extended lexicon word matching described in **4.**. The street number is assumed to be the first field of the street line. If ZIP code line includes only the ZIP code, the second preceding line is first assumed to be the street line, otherwise the immediate preceding line is assumed to be the street line. If the likelihood of the detected street number is less than a threshold, up to two preceding lines are assumed successively to be the street line until the street number with sufficient likelihood is detected. In actual word pre segmented images, street number fields often split and divided into several pieces, which have to be merged again into a field. This problem is resolved through multiple use of word recognition algorithm to a set of successive word segments.

2.2 Street Name Recognition

Street name is recognized by the extended lexicon word matching. The lexicon is generated through the ZIP+4 directory search for given pair of ZIP code and street number. The lexicon is first generated for top candidates of ZIP code and street number. If the generated lexicon is null, it is generated for second candidate of ZIP code and the top candidate of street number.

The street name recognition is performed in long word lexicon scheme, i.e. the pre directional, street name, and the suffix is concatenated in a word, and is dealt as a single word. The word images following the street number image are concatenated and supplied as a single word image to the word recognition algorithm.

2.3 Lexicon Generation

A basic (long word) street lexicon is a relation (set of records) which consists of the pre directional, street name and the suffix. The records, which match to a given pair of ZIP code and street number is collected from the ZIP+4 file. Other relations representing the abbreviation of the pre directional and the street suffix are predefined and used to expand the basic street lexicon in terms of the join operation (direct product). After the join operation, case variation (initially capitalized lower case character strings) is generated and added. For example a basic street lexicon:

- BAILEY AVE
- HARLEM RD
- MAIN ST
- MAPLE RD

is expanded to

- BAILEY -
- BAILEY AVE
- BAILEY AVENUE
- Bailey -
- Bailey Ave
- Bailey Avenue
- HARLEM - .....

where "-" denotes a null string. In this example pre directionals are all null string. Each record is concatenated in a single word preceding the word matching. In the evaluation test the average lexicon size is about 80 and 480 before and after the expansion respectively.

Fig.2 shows an output of each stage of street address recognition.

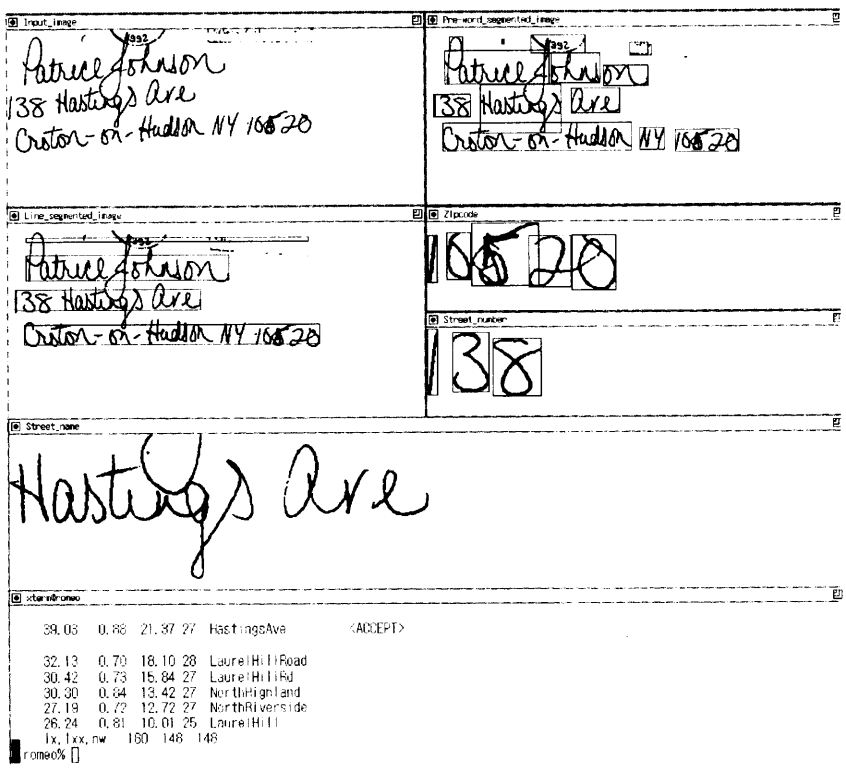

Fig.2 Output of each stage of street address recognition.

3. POBox Number Recognition Subsystem

3.1 POBox Number Location and Recognition

POBox number is recognized by the extended lexicon word matching. The POBox number is assumed to occupy entire POBox line with a preceding POBox key word such as "PO", "POB", "POBOX", "BOX", ... The rest of the POBox number location works in the same way as in the street number location.

The POBox line recognition is performed in the long word lexicon scheme, i.e. the POBox keyword and the POBox number are concatenated in a word, and is dealt as a single word. The word images in a POBox line are supplied as a single word image to the word recognition algorithm.

Fig.3 shows an output of each stage of POBox address recognition.

4. Extended Lexicon Word Matching

An extended lexicon word matching is an extension of a lexicon directed word matching [1],[2], which allows the lexicon to include lexicon words with variables (wild cards). In the lexicon directed algorithm, the ASCII lexicon of possible words is utilized. Given a lexicon word, the primitive character segments obtained by over segmentation of an input word image are merged and matched against a letter in the lexicon word so that the average character likelihood is maximized. The word matching process is repeated for all lexicon words and the lexicon words are sorted and ranked according to the average character likelihood. The extended lexicon algorithm is obtained from the lexicon directed algorithm by the following modification. The likelihood for the wild card is calculated by taking the maximum likelihood for all letters, and the associated letter is substituted to the wild card.

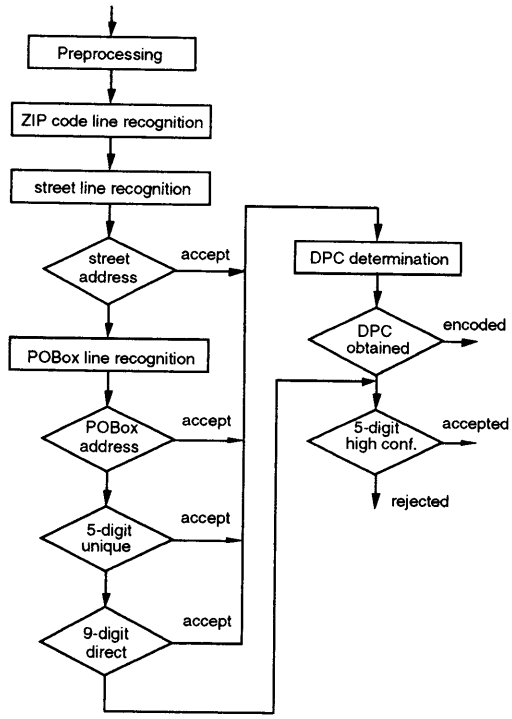

Fig.1 Block diagram for handwritten address interpretation.

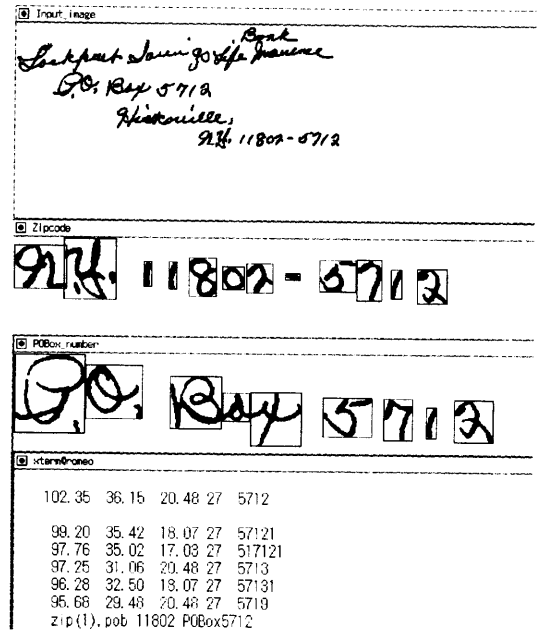

Fig.3 Output of each stage of POBox address recognition.

For example, it significantly simplifies the task of the POBox number recognition, if the lexicon can include such words as:

PO%%	POB%%	POBOX%%
PO%%%	POB%%%	POBOX%%%
PO%%%%	POB%%%%	...etc.
PO%%%%%	POB%%%%%	
PO%%%%%%	POB%%%%%%	

Where the '%' is a wildcard for numerals. This extended lexicon is highly advantageous because pre segmentation of POBox keyword and POBox number is not needed. Generally the pre segmentation of POBox keyword and POBox number is not easy task. It is possible to generate the smallest sufficient POBox lexicon after recognizing the ZIP code. However in the evaluation test, a POBox lexicon of fixed size (30) is used.

Besides the POBox recognition, it is also advantageous for ZIP code recognition to use extended lexicon words with two letter sate name prefix:

NY14%%%
NY14%%%-%%%%
MI48%%%
MI48%%%-%%%% ....etc.

In this case two letter state name and the first two digits of a ZIP code are recognized by more reliable lexicon directed (context sensitive) mode than lexicon free mode. These lexicon words enable the word matching to handle mis segmented state name and ZIP code. Pure numeral lexicon words

%%%%%
%%%%%-%%%%

for 5 digit and 9 digit ZIP code are also used together. In the evaluation tests an extended ZIP code lexicon of size 514 (including case variation) is used.

Street number recognition is performed using lexicon words:

%
%%
%%%
%%%%
%%%%%

Additional non-numeral street numbers derived from the zip+4 files, e.g. "One", "%%%-A", ... can be used, if necessary. However in the evaluation test, the street number lexicon of fixed size of 5 (one to five digit pure numeral strings) is used.

The extended lexicon word matching is convenient for secondary name recognition with secondary number, e.g. suite%%, suite%%%, APTS%%%, BLDG%%%, ... The extended lexicon word matching is flexible and sufficient for all address block word recognition tasks and is suitable for hardware implementation.

5. Determination of Address Type and DPC

5.1 Determination of Address Type

The address interpretation system assumes four address types, i.e. street address, POBox address, unique 5-digit ZIP address, and direct 9-digit ZIP (on mail piece) address. The type of address is determined in this order by sequential test (Fig.1). Once the address type is determined, the rest of the test is not applied, e.g. if it is accepted as street address, it is never accepted as POBox address or others.

The street line recognition subsystem generates several ranked pairs of street number and street name for given 5-digit ZIP code. If the top candidate pair is accepted the type of input address is determined to be street address. The top candidate pair is accepted if the ZIP code, street number and street name are all accepted. The ZIP code is accepted if the confidence (likelihood) of the first ranked ZIP code is greater than t_{z1}, and the difference between the first and the second ranked ZIP codes is greater than t_{z2}. The street number and street name are accepted or rejected in the same way using threshold pairs (t_{s1}, t_{s2}) and (t_{t1}, t_{t2}) respectively.

The POBox line recognition subsystem generates several ranked POBox number for given 5-digit ZIP code. If the top candidate is accepted the type of input address is determined to be POBox address. The top candidate is accepted if the ZIP code and POBox number are both accepted and the pair is in the ZIP+4 file. The POBox number is accepted or rejected in the same way using a threshold pair (t_{p1}, t_{p2}).

If the top ranked ZIP code is accepted and the record type is "14" in the City/State file, the type is determined to be unique 5-digit ZIP address. The top ranked ZIP code is accepted in the same way but with tighter thresholds (t_{z1}', t_{z2}') than for street and POBox address.

If the top ranked ZIP code is 9-digit and is accepted with the tighter thresholds, the type is determined to be direct 9-digit ZIP address.

If the top ranked ZIP code is 5-digit and is accepted with the tighter thresholds, the address is accepted. Otherwise it is rejected.

In the evaluation test the thresholds are fixed as follows.

$(t_{z1}, t_{z2}) = (-100, 0)$		; ZIP code
$(t_{s1}, t_{s2}) = (160, 5)$		; street number
$(t_{t1}, t_{t2}) = (40, 7)$	for 40% encode rate	
$(20, 5)$	for 50% encode rate	; street name
$(t_{p1}, t_{p2}) = (105, 3.5)$		; POBox number
$(t_{z1}', t_{z2}') = (50,6)$		; ZIP code (unique 5-digit, direct 9-digit, 5-digit accept).

The threshold pair (t_{z1}, t_{z2}) is temporally so loose that most of ZIP codes are accepted unconditionally. Because the length of ZIP code is known *a priori*, its confidence is relatively higher than street number and POBox number.

5.2 Determination of DPC

Delivery point code is determined separately for each address type.

For the unique 5-digit ZIP address and the direct 9-digit ZIP address, the ZIP code itself is the DPC.

For the street address, the ZIP code, street number and street name is encoded, and for the POBox address, the ZIP code and POBox number is encoded to the DPC using the ZIP+4 file.

For the street address, there can be multiple DPC's for given ZIP, street number, and street name. To select the finest possible depth of DPC the following algorithm was employed.

Algorithm; DPC selection of finest possible depth

1) If two DPC's are obtained and the record type of the one is street and the other is not street (e.g. high-rise, firm, ...), output the DPC of the non street record.
2) If three or more DPC's are obtained and there is only one street DPC, output the street DPC.
3) Otherwise reject.

5.3 DPF Post processing for Error Reduction

In current research system, the ZIP+4 File is used for street name lexicon generation, and the Delivery Point File (DPF) is used for post processing to detect and reduce the encoding error. The average size of street name lexicon generated from ZIP+4 is about 80, while it is about 20 if generated from the DPF. Therefor three fourth of street names in miss-encoded addresses are not in the DPF, i.e. the error rate is reduced to one fourth through the DPF search for obtained ZIP code, street number, and street name.

The ZIP+4 File has 28 million records, each of which contains ZIP code, 4digit add-on, record type, and street name information including pre- and post- directionals and the suffix. The DPF has a 100 million records, each of which correstponds to a mail stop, i.e. a house, an apartment, or a suite in a building.

6. Performance Evaluation

6.1 Performance Evaluation of Street Line Subsystem

Tables 1 and 2 show the result of street name recognition. The performance was evaluated using "bha" test samples. Among the samples from bha_6000 to bha_7603, 1330 street address samples (including high-rise, firm, 5-digit and reject) were used for this test.

Table 1 summarizes the cumulative correct rates of street name recognition. The top correct recognition rate was 62.18%. The top correct recognition rate of ZIP code and street number was 82.48% and 69.77% respectively, and the rate of correct pair was 62.18%.

Table 2 summarizes the tradeoff between reject and top-choice error rates at different operating points. For example, at (t_{f1}, t_{f2}) = (40.0, 7.0), the error rate was 2.03% with 59.17% rejection. In this table "correct" means all correct in recognition of ZIP code, street number and street name, while in the Table.5 "correct" means correct in recognition of street name but not necessary in street number or ZIP code.

Table 1. Cumulative Correct Rates of Street Name Recognition

Top-N choice	no.	%
1	827	62.18
2	843	63.38
5	850	63.91
Total	1330	100.00

Table 2. Error v.s. Reject of Street Name Recognition
(Correct: All correct in ZIP code, street number and street name)

t_{t1} t_{t2}	Reject	Error	Correct
10.0 3.0	45.11 (600)	9.32 (68)	90.68 (622)
20.0 5.0	49.47 (658)	5.06 (34)	94.94 (638)
40.0 7.0	59.17 (787)	2.03 (11)	97.97 (532)
50.0 9.0	70.53 (938)	1.53 (6)	98.47 (386)

6.2 Performance Evaluation of POBox Line Subsystem

Tables 3 and 4 show the result of POBox number recognition. The performance was evaluated using "bha" test samples. Among the samples from bha_6000 to bha_7603, 204 samples having POBox line were used for this test.

Table 3 summarizes the cumulative correct rates of POBox number recognition. The top correct recognition rate was 70.94%. The top correct recognition rate of ZIP code was 84.73%, and the rate of correct pair was 61.58%.

Table 4 summarizes the tradeoff between reject and top-choice error rates at different operating points. For example, at (t_{p1}, t_{p2}) = (90.0, 3.5), the error rate was 1.02% with 51.96% rejection. In this table "correct" means both correct in recognition of ZIP code, and POBox number, while in the Table.3 "correct" means correct in recognition of POBox number but not necessary in ZIP code.

Table 3. Cumulative Correct Rates of POBox Number Recognition

Top-N choice	no.	%
1	144	70.94
2	150	73.89
5	155	76.35
Total	203	100.00

Table 4. Error v.s. Reject of POBox Number Recognition
(Correct: Both correct in ZIP code, and POBox number)

t_{p1} t_{p2}	Reject	Error	Correct
0.0 0.5	34.80 (71)	9.77 (13)	90.23 (120)
85.0 1.0	39.22 (80)	6.45 (8)	93.55 (116)
80.0 3.0	44.12 (90)	4.39 (5)	95.61 (109)
90.0 3.5	51.96 (106)	1.02 (1)	98.98 (97)

6.3 Performance Evaluation of Integrated System

Tables 5 and 6 show the error v.s. encode rate of the integrated system. The performance was evaluated using "bha" test samples. All the samples from bha_6000 to bha_7603 were used for this test.

Table 5 and 6 summarizes the performance with and without DPF post processing at different operating points, respectively. The error rate is 1.12% with 50.19% encode rate, and 0.87% with 43.12% encode rate when error reduction is performed through DPF search. The error rate was reduced from 4.11% to 1.12%, from 2.15% to 0.87% respectively.

Table 5. Error v.s. Encode rate with DPF post processing

t_{t1} t_{t2}	Encode rate	Error	Correct
20.0 5.0	50.19 (803)	1.12 (9)	98.88 (794)
40.0 7.0	43.12 (690)	0.87 (6)	99.13 (684)

Table 6. Error v.s. Encode rate without DPF post processing

t_{f1} t_{f2}	Encode rate	Error	Correct
20.0 5.0	51.75 (828)	4.11 (34)	95.89 (794)
40.0 7.0	43.69 (699)	2.15 (15)	97.85 (684)

7. Summary and Future Works

In this paper, integrated address interpretation system using an extended lexicon word matching was described.

In the evaluation test for 300dpi address block images, the error rate was 1.12% with 50.19% encode rate when DPF was used for error reduction post processing. This result is significantly better than the result ever reported by other investigators[4], and is sufficiently better than the target, which is 50% encode rate with 2% error rate.

If DPF is directly used in street name lexicon generation, the performance will be further improved. The performance of ZIP code recognition will be also improved if City/State name recognition subsystem is integrated. Rural route recognition subsystem is not implemented yet. All rural route addresses were rejected except one, which was miss recognized to a POBox address. All foreign mails and mails rejected by human truther were also rejected by the system.

Recognition speed is currently about 10sec/address on SPARC Station 20. Program optimization will improve the recognition speed considerably.

Acknowledgments

The work reported in this paper was supported in part by a contract from USPS(RFP1044230-91-A-0036). The authors would like to acknowledge Dr. B. Phan of USPS and Dr. J. Tan of Arthur D. Little for their comments and useful suggestions. The authors also would like to acknowledge SUNY Buffalo for their effort in providing us with address block images for this work.

References

[1] F.Kimura, M.Shridhar, and Z.Chen, "Improvements of a Lexicon Directed Algorithm for Recognition of Unconstrained Handwritten Words", Proc. of 2nd ICDAR, pp.18-22 (1993).

[2] F.Kimura, S.Tsuruoka, Y.Miyake, and M.Shridhar, "A Lexicon Directed Algorithm for Recognition of Unconstrained Handwritten Words", IEICE trans. on Inf. & Syst., Vol.E77-D, No.7, pp785-793 (1994).

[3]F.Kimura, Y.Miyake, and M.Shridhar, "Handwritten ZIP Code Recognition Using Lexicon Free Word Recognition Algorithm", Proc. of 3rd ICDAR, pp.906-910 (1995).

[4]S.N.Srihari, V.Govindaraju, and A.Shekhawat, "Interpretation of Handwritten Address in US Mailstream", Proc. of 2nd ICDAR, pp.291-294 (1993).