手書き宛名画像を対象とした文書理解システムの改良

森本　隼人*　・　岡村　健史郎**　・　田中　裕貴***　・　末弘　光次郎****

Improvements of Document Understanding System for Handwritten Address Image

Hayato MORIMOTO* , Kenshiro OKAMURA** , Yuuki TANAKA***
and Koujiro SUEHIRO****

Abstract

This paper proposes a method of document understanding system for handwritten address image. In a previous work, the performance of rough classification is low, and final outputs are not so accurate. The causes of these problems are low rate of character recognition and fault of knowledge-based processing. To avoid these problems, the proposed method uses multiple candidates for the character recognition and reselects most available one by using address phrases. Furthermore, by changing the priority in rough classification and giving the dynamical threshold in rough classification, the method can overcome the fault of the knowledge-based processing. In experimental results, the effectiveness of our method has been proven.

Key words: document understanding system, address reading, character segmentation

1. まえがき

大量の郵便物が世の中を行き交う現在において、郵便物の仕分けの効率化は重要な問題である。現在では郵便番号による仕分けが行われているが、より効率よく仕分けを行うために文書理解システムを用いる方法が注目されている。文書理解システムによる仕分けは、文字認識技術を用い、宛名画像に書かれている宛名をそのまま読み取るものである。この方法は郵便番号を用いた仕分けよりも誤認識に強く、より詳細な仕分けが可能という利点がある。しかし、現在、文書理解システムによる仕分けは実用化には至っておらず、実現に向けて、住所データベースを用いて文字認識誤りを補う方法や、仮説生成と検証を組み合わせた方法[1]など、さまざまな方法が研究されている。

筆者らが研究を進めている文書理解システムは、手書き宛名画像から文字列を生成するパターン処理部と、蓄えた知識を基に知識処理を行う知識処理部の二つから構成されている。この従来のシステムのパターン処理部は、画素密度を用いた文字切り出しを行っており、文字切り出し率が高く、縦書き、横書きどちらに対しても事前知識なしに適応できる利点をもっている[2]。一方、個々の文字の認識率が低く、結果として住所検索率が低くなるという欠点があった。これを補うために、認識結果の下位候補を用いる方法が考えられる。下位候補まで用いることによって、見かけ上の認識率を上げることが可能であるが、認識文字から作る住所の組み合わせが莫大な量となり、処理時間に大きな影響を与えてしまう。

そこで、下位候補を含む文字認識候補から認識文字として最も確からしい文字を、大分類[3]の結果を用い、新たに選択する方法をとった。これにより、処理時間に影響を与えることなく文字認識率をあげることできるようになった。一方、従来システムにおける知識処理部には、大分類の精度が低く、正解住所が大分類の結果に含まれないという問題点があった。これを改善するために、大分類によって住所を絞り込む際の基準となる値の定義と、大分類の絞込みの範囲を変更した。本論文では、いくつかのシミュレーション実験の結果から、これらの手法の有効性を明らかにする。

*電子･情報システム学専攻　**情報工学科　***豊橋技術大学　****綜合警備保障株式会社

2005年 9月30日 受付

2. 従来システムの構成と問題点

2.1 処理の流れ

ここで、本研究で開発したシステムの処理の流れを解説する。図1に示すように、手書き宛名画像を入力データとし、文字切り出し、文字認識処理からなるパターン処理部によって、認識文字列を得る。その後、大分類、距離計算からなる知識処理部によって、文字認識結果と一番近い住所を検索結果として出力する。

2.2 各処理の詳細

2.2.1 文字切り出し

本システムでは、画素密度による文字切り出しを行っている。図2に縦書き宛名画像とその文字切し結果を、図3に横書き宛名画像とその文字切り出し結果をそれぞれ示す。図2(b)、図3(b)の、小さい領域を持つ枠は各切り出し文字を、複数の切り出し文字を囲んでいる枠は文字列を意味する。

2.2.2 文字認識処理

文字切り出しを行った個々の文字に対しての文字認識処理は、加重方向ヒストグラムとマハラノビス距離を用いる[4]。表1に示すように、従来システムの文字認識率は縦40%、横60%と低い。これは縦書き文字画像の文字認識結果が10文字の場合、4文字しか正しく認識されてない事を意味する。この原因として宛名を書く際にくずし字で書き、字が大きく変形してしまう事などがあげられる。文字認識率の低さは検索率の低さに直接影響する問題であり、最も改善が要求される問題点である。この問題を改善するため、下位候補を含む文字認識候補から文字を選択し直し、新たに文字列を作成する手法に変更をした。

2.2.3 大分類

大分類とは、全国12万件の住所データベースから、パターン処理部が生成した文字認識結果と類似した住所を絞りこむ処理のことである。これは、12万件の住所に対して距離計算を行ったのでは莫大な処理時間がかかってしまうことを改善するために用いる処理である。この大分類に用いる文字一致度の求め方を図4に示した。

大分類では、住所データベースを本システム用に変換し、認識対象となる約4000種類の文字が各住所の何文字目に存在するのかという情報を文字ごとに整理した文字データベースを事前に作成する。これを用いて認識文字列に含まれる個々の文字がどの住所に含まれているかを検索し、該当住所の文字一致度を一つ増やす。図4を例にとって具体的に説明する。まず、認識文字列中の「札」がどの住所に含まれているかを文字データベースを参照し調べる。これによって「札」は住所番号1、2、3の住所に含まれていることが分かる。こ

表1　文字切り出し率と認識率

	文字切り出し率	文字認識率
縦書き	91.0%	40.0%
横書き	89.0%	60.0%
平均	90.0%	50.0%

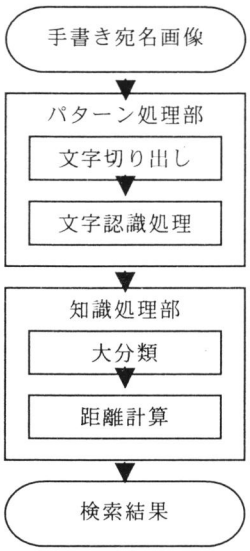
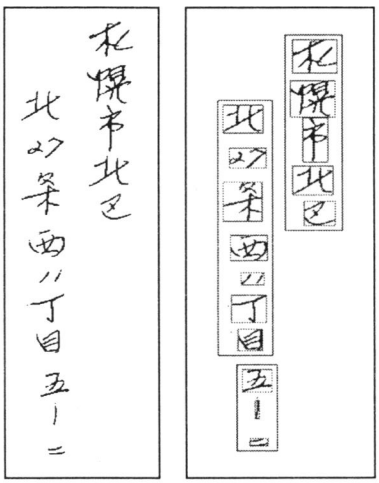
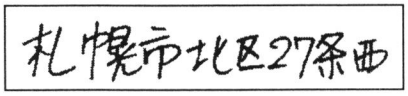


図1　処理の流れ

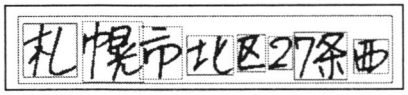


(a)　手書き宛名画像　(b)　文字切り出し結果

図2　縦書き宛名画像



(a) 手書き宛名画像



(b) 文字切り出し結果

図3　横書き宛名画像

こで得た住所の文字一致度にそれぞれ 1 加える。

この処理を認識文字列の全文字に対して行うことで求まった最終的な文字一致度から優先度を求める。この優先度が最大のものから数えて N 件目までの住所を候補住所とする。

・　優先度に関する問題点

優先度とは、大分類によって住所データベースから候補住所を絞り込む際に基準となる値で、式(1)で定義される。

$$R_i = \frac{m_i}{M_i} \quad \cdots(1)$$

ここで認識文字列に対する住所 i の優先度を R_i、文字数を M_i、文字一致度を m_i とする。

リスト 1 に認識文字列の例、候補住所の例をリスト 2 に示す。リストに示した認識文字列と候補住所の一行目を比較すると、認識文字列は住所データベースに存在する住所である「北区」を含んでいることが分かる。よって式(1)より、住所データベース中の「北区」の優先度 R_i は最大値の 1.0 となる。このように、文字長が短い住所の方が優先度が高くなりやすく、その結果誤った住所が候補住所に選ばれてしまう。これらを改善するために、優先度の式を変更した。

・　絞込み件数に関する問題点

従来にシステムでは、大分類によって候補住所を絞り込む件数を優先度が高い住所から N 件と定数に設定していた。これでは、N 件目の優先度と、N+1 件目の優先度が等しかった場合、N 件目と N+1 件目の重要さは同じにもかかわらず N+1 件目は距離計算の対象からはずされることになる。これを改善するために、絞込み件数を定数ではなく、動的に決定する手法に変更した。

・　文字一致度に関する問題点

リスト 3 に大分類に失敗した認識文字列の例を、リスト 4 に候補住所の例をそれぞれ示す。候補住所例の二行目をみると、認識文字列と一致した文字の並びは、川、市、西となっている。しかし、認識文字列をみると文字の並びは市、川、西となっている。このように文字の並びを考慮せずに文字一致度を求めてしまうと、関係ない住所が候補住所に含まれてしまう。これを改善するため大分類の、文字一致、不一致の判断に変更を加えた。

2.2.4　距離計算

距離計算とは、大分類が絞り込んだ候補住所から認識文字列に最も近い住所を動的計画法によって検索し、これを検索結果とする処理である[5]。

3.　文字認識率と大分類成功率の向上

2.3 で述べた問題点の改善方法について詳しく解説する。

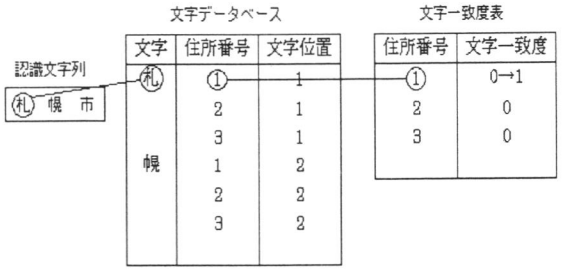


図 4　文字一致度の求め方

札幌常北区汝岩四簾両

リスト 1　認識文字列

北区　　一致度: 2 優先度:1.000
札幌市北区　一致度: 4 優先度:0.800

リスト 2　候補住所

ふ清市ふ川西町

リスト 3　大分類に失敗した認識文字列

川西市栄町　一致度: 4 優先度:0.800
桶川市西　　一致度: 3 優先度:0.750

リスト 4　候補住所

3.1　パターン処理部

2.3.1 で述べたように、パターン処理部には文字認識率が低いという問題点がある。これを改善するために、下位を含む文字認識結果を用いる手法をとった。以下、下位を含む文字認識結果を文字認識候補と呼ぶ。従来システムでは、文字認識処理によって最も特徴が近いと判断された文字を 1 位の認識文字列として利用し、それ以外の文字認識結果は利用されていなかった。しかし、2 位以降の文字にも正解文字が含まれている可能性は十分に有る。この 2 位以降の文字認識結果を下位候補と呼び、これを利用することによって全体的な文字認識率の上昇が期待できる。しかし、文字認識候補を用いると、これらの組み合わせで作る文字列の数が莫大な量となり、処理時間に影響を与えてしまう。

それを防ぐために、文字認識候補の中から、最も確からしい文字を選択する。具体的には、文字認識候補の i 番目の切り出し文字に対する k 番目の認識結果を c(i,k)として、認識結果 c(i,1)、c(i,2),…, c(i,k),…, c(i,K)の中から最も確からしい文字を選択し直し、これを再利用する手法を

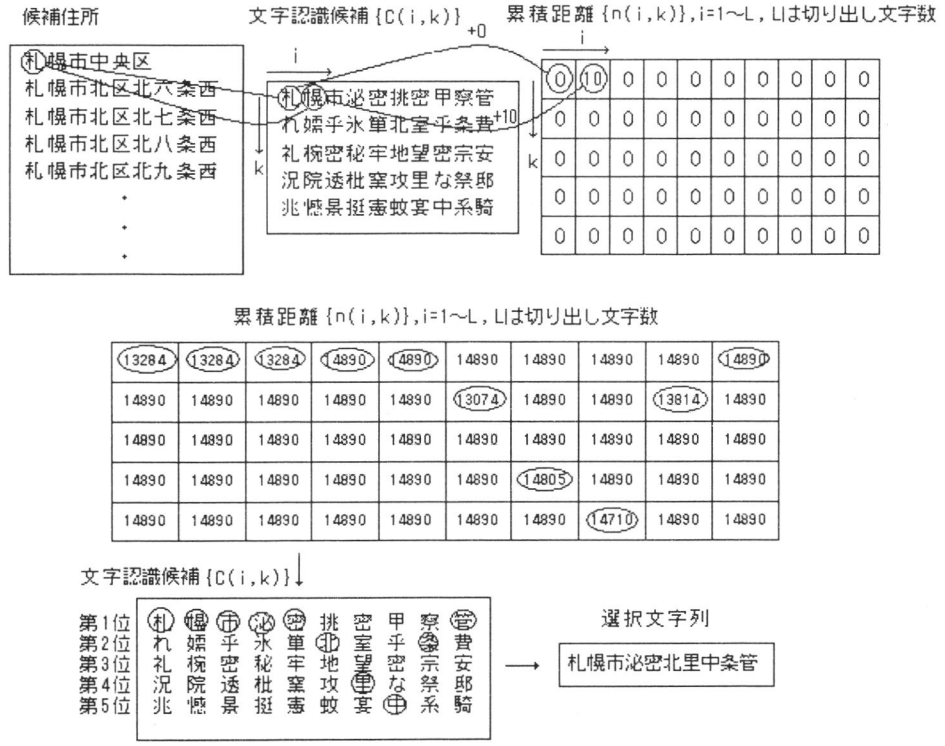


図5　文字選択の処理図

とった。この選択は大分類の結果である候補住所を用いて行う。

まず認識候補中の文字 c(i,k)に対する確からしさを表す累積距離 n(i,k)を定義する。文字認識候補{C(i,k)}中のある文字を c(i',k')とした時に、候補住所中の文字と c(i',k')が一致した場合は、対応する累積距離 n(i',k')に候補住所の切り出し位置と i'の差を加える。候補住所の文字が c(i',k')と一致しなかった場合は、候補住所の文字数と認識候補の文字数のうち、その大きい方を加える。この処理を候補住所中の全文字に対して行う。この具体例を図7に示す。図7は、候補住所中の「札」という文字が、c(1,1)と一致したが、文字切り出し位置と i の差が0であるため n(1,1)には値を加えず、c(1,2)に対しては一致しなかったため、認識候補の文字数である 10 を n(1,2)に加えたという例である。図7に示すように、最終的に{n(i,k)},k=1～K の各 i の中から、値が一番小さかった文字を選択する。値が同じだった場合は、k が小さいものを選択する。そうして得られた各文字を並べたものを選択文字列として出力する。これは、別の基準で文字認識を再び行うことを意味する。

3.2　知識処理部

3.2.1　優先度の変更

2.3.2 で述べたように、従来システムの優先度の定義には、文字数が少ない住所が候補住所に選ばれやすいという問題が存在する。それを改善するために、式(1)を式(2)のように変更した。

$$R_i = \frac{m_i}{M_i'} \quad \cdots(2)$$

式(1)では右辺の分母を、優先度を求める住所 i の文字数としていたが、式(2)では右辺の分母を認識結果の文字数とした。これによって、文字数が少ない住所の優先度が高くなることを防ぐことができる。

3.2.2　絞込み件数の変更

2.3.2 で述べたように、絞込み件数を定数で設定すると、必要な情報まで切り捨ててしまう可能性がある。これを改善するために、N 件目の優先度を基準とし、優先度が基準以上であった場合はその住所を候補住所に含めるという方法に変更した。これによって、これまで利用されていなかった住所を候補住所として用いることができるようになる。

3.2.3　文字位置情報の追加

大分類の処理において、候補住所中の文字と認識結果中の文字が一致した場合でも、文字位置が二文字分以上ずれていれば、その場合は不一致として処理する。さらに式(2)右辺の分子 m_i を、文字位置情報を加味して求めるように大分類を変更することにより、従来よりも高精度に候補住所を

選びせるようになることが期待できる。

4. シミュレーション実験

3 章で示した改善方法をシステム上に実現しその有用性をシミュレーション実験によって確認した。表 2 に各手法を用いた縦書き宛名画像を対象とした実験結果、表 3 に横書き宛名画像を対象とした実験結果を、表 4 にこれらの平均をそれぞれ示す。なお、表 2, 3, 4 において、

①従来システム
②文字認識候補を用いたシステム
③優先度の定義を変更したシステム
④絞り込み件数を変更したシステム
⑤文字位置情報を追加したシステム

とし、表中の「+」は、「+」でつながれたシステムそれぞれを合成したことを意味する。

4.1 実験に用いたデータ

本システムで用いるデータは、縦書き宛名画像、横書き宛名画像、そして 121637 件の住所データベースである。縦書き宛名画像は、郵政総合研究所が作成した手書き漢字画像データベース IPTP CD-ROM2 を、横書き宛名画像は、大島商船高専の学生が作成した物をそれぞれ 340 件使用した。住所データベースは、郵政省が作成した全国 47 都道府県の住所データを本システム用に変換したものを使用した。

4.2 文字認識候補を用いた実験

3.1 で解説した手法を従来システムに追加し、文字認識候補をもちいた実験を行った。縦、横書き宛名画像に対してパターン認識を行った 5 位までの文字認識候補を入力データとしてシミュレーション実験を行った。表 2 の①+②に示した結果をみると、検索成功率、大分類成功率共に 5〜10%の精度向上が見られた。このことから、これまで利用していなかった下位候補に正解文字が多く含まれていたことが分かる。しかし、表 3 の①+②に示した結果をみると、大幅な変化はなかった。これは、横書きの文字認識率が、縦書きの文字認識率に対して比較的高かったことが原因だと考えられる。そのため、文字認識候補を用いた場合でも認識率に大きな変化が無かったと考えられる。

これらの事から、文字認識率が低い場合、文字認識において第 1 位となった文字だけではなく、認識結果の下位候補を用いることは有効な手法であると考えられる。

4.3 優先度変更後の検証実験

2.3.2 で解説した優先度に関する問題点を改善するために、式(1)を式(2)に変更し、優先度変更が、大分類の精度どのような影響を与えたかをシミュレーション実験によって確認する。表 2、表 3 の①+②+③に示したように、縦横共に精度の向上が見られた。これは、3.1 で述べたように、文字数が少なく、候補住所として不適切な住所が候補住所に含まれてしまう問題が改善されたためであると考えられる。リスト 5、リスト 6 に 3.1 で示した問題点が表れている候補住所と、優先度の改善によって得られた候補住所をそれぞれ示す。リスト 5、リスト 6 を比較すると、変更前では文字数が少なく、正解ではない住所が候補住所に含まれていたが、変更後ではそれが改善されていることが分かる。

4.4 絞込み件数変更後の検証実験

3.2.2 で解説した絞込み件数変更が、大分類の精度にどのような影響を与えたかをシミュレーション実験によって確認する。表 2、表 3 の①+②+③+④を見ると大分類成功率は高くなっているの

表 2　縦書き結果

	①	①+②	①+②+③	①+②+③+④	①+②+③+④+⑤
検索成功率	40.0%	49.1%	54.7%	53.0%	65.2%
大分類成功	43.3%	58.2%	61.2%	65.6%	84.4%

表 3　横書き結果

	①	①+②	①+②+③	①+②+③+④	①+②+③+④+⑤
検索成功率	70.0%	67.3%	70.3%	69.1%	72.6%
大分類成功	80.0%	75.3%	76.2%	80.0%	88.8%

表 4　縦横平均

	①	①+②	①+②+③	①+②+③+④	①+②+③+④+⑤
検索成功率	55.0%	58.2%	62.5%	61.5%	68.9%
大分類成功	61.7%	66.8%	68.7%	72.8%	86.6%

日田市隈	一致度:3 優先度:0.750
日田市	一致度:2 優先度:0.667
島田市	一致度:2 優先度:0.667
島田市元島田	一致度:4 優先度:0.667
日田市北友田	一致度:4 優先度:0.667

リスト5 優先度改善前の候補住所例

常呂郡常呂町日吉	一致度:5 優先度:0.500
常呂郡常呂町北進町	一致度:5 優先度:0.500
海部郡飛島村飛島新田	一致度:5 優先度:0.500
広島市安佐北区口田南	一致度:4 優先度:0.400
札幌市北区北三四条西	一致度:4 優先度:0.400

リスト6 優先度改善後の候補住所例

に対して、検索成功率が低下する傾向が見られる。これは、この変更によって絞込件数が大幅に増加したことによる、距離計算の誤検索が増加したことが原因だと考えられる。

4.5 文字位置情報を追加したシステムの実験

3.2.3 で解説した、住所データベース中の住所と、文字認識候補間の文字一致の判定に文字位置情報を加味するように変更を加えたシステムのシミュレーション実験を行った。表2、表3の①+②+③+④+⑤を見ると、変更後に知識処理の精度が高くなっていることが分かる。これは、文字認識候補と関連が薄い住所が候補住所に含まれにくくなったためであると考えられる。

4.6 今後の課題

以上のシミュレーション実験から得た情報を基に考察すると、大分類成功率 86.6%と、従来システムと比べて 24.9%向上した。

しかし、大分類に失敗した文字認識候補を解析した結果、一文字も認識成功していなかったものが 61.4%、市町村などの特徴の無い文字しか認識成功していなかったものが 27.3%と、これら大分類に成功させることが不可能といって良いような文字認識候補が、失敗したデータの 88.7%を占めていた。この事から、大分類による絞込みは限界に近づきつつあることが分かる。そのため今後の課題としては、文字認識率の向上を図ることと、検索成功率を向上させるために距離計算部分に変更を加えていくことが挙げられる。さらに、従来システムのような、パターン処理から知識処理への一方向の処理ではなく、優先度などの値をキーとした、パターン処理部のフィードバック処理も検討する必要があるといえる。

5. まとめ

本論文では、郵便物の仕分けにおける文書理解システムについて論じた。従来システムの問題点である、文字認識率の低さ、優先度の定義、絞り込み範囲の改善手法を提唱し、その手法によって検索率の向上が可能であることを示した。

今後は、文字認識率の向上や、フィードバック処理の追加を検討し、さらに、本システムで得た情報を基に、郵便物の配達経路などを設定するなど、別システムとの連携や、宛名画像だけではなく、書類などの一般文書画像への応用も視野に入れ、文書理解システムの広い分野での応用を目指す。

謝辞

実験に使用した ETL9B 文字データベースを提供して下さった産業総合技術研究所、そして縦書き宛名画像に関して手書き漢字データベース IPTP CD-ROM を提供して下さった郵政総合研究所に感謝致します。

参考文献

[1] 下村秀樹, 福島俊一, 山内俊史 : 仮説生成と検証の効率的組合せに基づく手書き文字列読み取り向け知識処理方式、情報処理学会論文誌、Vol.40 No.7 pp.2905-2917、1999

[2] 岡村健史郎, ユジン・クルズ, 佐長康久, 浜本義彦 : 画素密度検出エージェントによる文字列の抽出と文字切り出し、大島商船高等専門学校紀要 第36号、2003

[3] 徳本一崇, 鈴木雅人, 加藤寧, 根元義章 : 候補あて名の優先度付けによる高速大分類法を用いた手書きあて名認識システム、電子情報通信学会論文誌 D-II、Vol.J84-D-II No.1 pp.83-92、 2001

[4] 田中裕貴 : 手書き文字画像を対象にした文字認識システムの作成、大島商船高等専門学校卒業論文、2005

[5] 柳沙織里 : 手書き宛名画像を対象にした住所認識システムの作成、大島商船高等専門学校卒業論文、2004