卒業論文

文字認識アルゴリズムに関する基礎的研究

東北大学工学部 電気・情報系 阿曽研究室

金津知俊

１９９３年３月

目 次

1 序論 **1**

1.1 本研究の背景及び目的 1

1.2 本研究の概要 2

1.2.1 認識対象文書 2

1.2.2 認識システムの作成と高精度化 2

1.3 本論文の構成 4

2 分離文字統合を行なう文書認識システム **5**

2.1 まえがき 5

2.2 文書認識システムの構成 5

2.2.1 スキャナ読み取り～行切り出し処理部 5

2.2.2 切り出し図形系列作成部 5

2.2.3 前処理、特徴ベクトル化部 7

2.2.4 ベクトルマッチング部 8

2.2.5 評価・統合部 8

2.3 認識実験 (1) 12

2.4 まとめ 14

3 部分パターン辞書の利用 **15**

3.1 まえがき 15

3.2 部分パターン辞書の作成 (横分離文字部分パターン辞書) 15

3.2.1 分離文字の抽出 15

3.2.2 部分パターンのベクトル化 17

3.2.3 部分パターン辞書のクラスタリング 17

3.2.4 部分パターン辞書の完成 19

3.3 評価部への導入 20

3.3.1 切り出し図形スコア評価部の改良 20

3.3.2 評価スコアへの加点方法 21

3.4 認識実験 (2) 23

3.5 まとめ 25

4 結論 **26**

4.1 本研究のまとめ 26

4.2 今後の課題 26

謝辞 **28**

参考文献 **29**

図目次

1.1 文書認識システムの構成 3

2.1 系統2の評価パスグラフと評価関数 9
2.2 norm-distance (8pt 明朝体) 10
2.3 norm-distance (12pt 明朝体) 10
2.4 norm-distance (20pt 明朝体) 11
2.5 norm-distance (12pt ゴシック体) 11

3.1 レーザプリンタ文字中の横分離文字数 16
3.2 明朝体部分パターン辞書のクラスタリング、グルーピングテスト 18
3.3 加点の例 (仮統合レベル1の場合) 22

表目次

2.1 認識実験結果 (1) 13

3.1 分離文字抽出サンプルフォント一覧およびその略称 16
3.2 クラスタリング、グルーピング用クラスター距離 19
3.3 認識実験結果 (2) 23

第 1 章

序論

1.1 本研究の背景及び目的

文書認識、特に日本語文字認識において文字切り出しの誤りは認識率の低下の大きな要因となり、切り出しの高精度化は重要な問題である。この研究は当研究室で開発された方向線素特徴量[1] を用いた文書認識システムの一部として、切り出しの高精度化を目指すものである。

ここで言う文字切り出しとは、光学式スキャナによりイメージとして取り込まれた文書画像から個々の文字を正しく抽出する作業を指す。しかし本研究では文字列をなす図形の集合が行として正しく分離されているものとし、この行イメージから１文字１文字を正しく切り出す作業を指すこととする。

行イメージから文字列を切り出そうとする場合、一般にヒストグラムの切れ目を用い画素の塊りを一文字として分離する方法が簡単で良く用いられる。しかし対象を日本語文書とした場合数多く存在する分離文字のためこれを正しく統合する処理が必要となる。分離文字というのは１文字中でも x,y 方向の射影に切れ目があるような文字で、行方向にこの切れ目がある場合問題となる (横書き時の「は」、「慣」、縦書き時の「う」、「言」など)。したがって、文字切り出しの高精度化とはこういった分離文字を正しく統合する方法を作成することに置きかえられる。

この統合に用いられる情報としては、従来よりピッチ情報が主として用いられてきた。これは日本語印刷文書の場合構成文字のほとんどがいわゆる「全角文字」であり、文字ピッチは行中でほぼ一定であるという仮定のもとでは最も有効な手段である。

しかし横書き文書では半角文字の混入でピッチが局所的に変化することが多く、とくに分離文字が隣接する場合、またトータルの文字数が少なく統計的情報が得られない場合など、更に高性能の製版システムが普及したことにより任意ピッチの文書が増えてきたことなど、ピッチ情報のみではもはや正しく切り出しを行うのは困難となっている。

そこでこの分離文字統合の高精度化のために認識結果を利用する方法が良く用いられる。ピッチ的に統合可能な複数の文字候補パターンに対し、イメージ上で統合した場合の認識結果とそれぞれのパターンに対する認識結果を比較し、統合、非統合を決定するものである。この方法によればピッチ自由度の高い文書においても高い精度で文字切り出しが可能であることが報告されている。

そこで本研究では、当研究室で考案され、個々の文字に対し高速高精度の認識が可能である方向線素特徴量を認識部に用い、その評価値を考慮した統合法を用いて高精度の切り出しを可能にすることを目的とする。

1.2 本研究の概要

1.2.1 認識対象文書

本研究で取り扱う実験用文書は、日本語 TEX により整形された横書き文書を 300dpi のプリンタで打ちだしたもので、文書中には半角の英数字も含まれる。この文書の特徴としては、ワープロ文書等に比べ文字ピッチがふぞろいであること、特に半角文字や記号などは各々固有の文字幅を持っているため、その混入によって文字ピッチが著しく変化することなどがあげられ、そのためピッチ情報のみでは正しい切り出しを行うことは不可能と考えられる。又他の特徴としては、印刷精度が高いのでつぶれや接触文字が少なく、認識が高精度に行えるということがあげられる。

1.2.2 認識システムの作成と高精度化

スキャナにより文書を読み取り認識結果として文字列を出力する一連の認識システムを図 1.1のように構成する。本研究でメインとなるのは図中の切り出し図形系列作成部と評価・統合部であり、高精度の分離文字統合のためのアルゴリズムを検討する。この切り出し図形系列生成部ではヒストグラムにより切り出された図形とともに、統合可能性のある隣接図形をイメージ上で統合した図形もすべて加えた切り出し図形系列を出力する。評価・統合部ではそれらすべてのマッチング結果から、選択的なものについては比較し統合するかどうかを決定、認識結果として文字列を出力する。この方法ではすべての可能な組み合わせについて切り出し図形とするためにマッチングの回数が増えるという短所もあるが、処理にフィードバックを含まないのでデータの流れを単純に出来る。またマッチングに使用する辞書としては通常の文字認識に使用される全角・半角文字のベクトル辞書を用い、認識実験により評価・統合部の基本性能を評価、以後の研究の基幹システムとして確立する。次に同システムに統合のための情報を追加して統合の高精度化する手段について検討する。本研究で検討したのは、分離文字の部分パターンを登録した部分パターン辞書をベクトルマッチング部の辞書に付加し、そのマッチング結果を用いて分離文字の統合を強化する手法である。

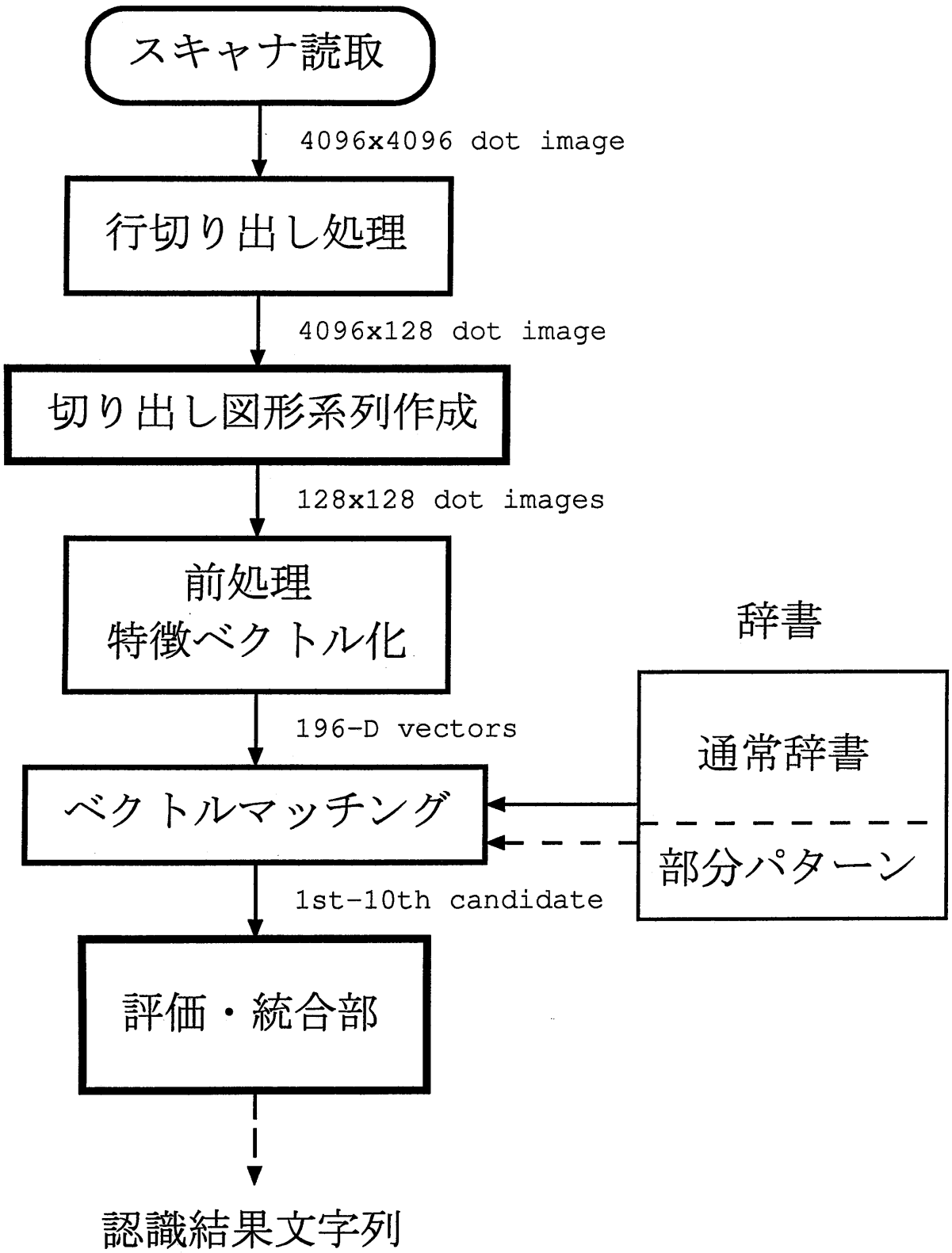


図 1.1: 文書認識システムの構成

1.3　本論文の構成

第１章 序論であり、本研究の背景と目的、および概要である。

第２章 第２章では前図1.1について各部を順に説明し一連のシステムを構築する。中でも最も問題となるのは、方向線素特徴ベクトルのマッチング結果から統合に利用する際の評価値の算出のしかたであり、特に分離文字の部分パターンなど構造の単純な図形が統合評価のとき問題となることを指摘し、これを改善するために特徴ベクトルの改良を含めた評価法を提案する。最後にこのシステムを使って文書認識実験を行い、結果をもとにこの単純比較法の限界について考え、これを基本システムとして付加情報による高精度化について述べる。

第３章 第２章で得られた基本システムの高精度化の一案として、部分パターンの利用について検討する。認識文字に対する辞書 (通常辞書) に分離文字の部分パターン辞書を追加しマッチング結果に反映することで統合の判断を確実にする方法を提案する。この部分パターン辞書作成に対する問題点及び、評価・統合部への導入の方法について考察し、実験結果より前手法からの改善効果を検討する。

第４章 全体の結論である。

第 2 章

分離文字統合を行なう文書認識システム

2.1 まえがき

本章では、当研究室で開発された方向線素特徴量を用い、認識結果を利用して切り出しの高精度化を行う文書認識システムを構成しその問題点を検討する。マッチングの辞書には通常の文字認識に使用する辞書(通常辞書：ＪＩＳ第１水準の漢字、仮名、英数字、記号等3303字種＋半角英数字62字種)のみを使う。統合部ではこのマッチング結果のみから評価値を求め、単純比較によって統合・非統合を決定する。この手法で高精度化を検討し、また他の評価材料を付加し更に確実な統合をおこなえるような基幹としての統合部を構築する。

2.2 文書認識システムの構成

2.2.1 スキャナ読み取り～行切り出し処理部

A4の横書き文書を市販の光学式スキャナで２値画像(4096 × 4096 dot)にして入力する。この画像データに対し傾き補正を行なった後,行方向の射影をもとに１行１行を分離する。但しここでは対象文書は数式等を含まないプレーンなテキストに限っているので行イメージデータ中には正しく１行が切り出されているとする。またこれらの作業は当研究室既存のシステムを用いるので本研究では特に扱わない。

2.2.2 切り出し図形系列作成部

前段で得られた帯状の行イメージから文字候補図形である切り出し図形を抽出する。
ここで本システム中唯一の仮定をおく

仮定１ 文字の基本スケールは縦横比１：１であり、１行中で文字ポイントは一定である。

この仮定を基にすると行中の文字集合に対し、ある定数として基本幅Wを設定することができる。この基本幅を用いて個々の文字の相対的な大きさが推測される。

行イメージから文字候補図形であるイメージ画素の塊りを抽出する。ここで行イメージ中には正しく１行が含まれており縦方向に複数の文字が存在しないこと、また縦方向の文字の重なりもないと考えるので、この作業は行イメージの列ヒストグラムの０値をもって画素を切り出すこと

によっておこなう。そのようにして切り出された各々横方向に独立な図形を切り出し図形と呼び、左のものから順にマップ番号をあたえ $c_0, c_1, \ldots, c_m$ とする。これらの集合を C とする。

$$C = \{c_0, c_1, c_2, \ldots, c_m\}$$

この切り出し図形集合 C に対し基本幅 W_c を次のように定める。

$$W_c = \max_{c_j \in C}(h_j)$$

h_j : 切り出し図形 c_j の高さ

かようにして得られた基本幅 W_c と、前述の仮定「縦横比１：１で行中の大きさは一定」を組みあわせ次の推測を得る。

推測１　行中の文字の横幅の最大値は W_c を越えない

この推測を基にすると、集合 C 中で隣接図形とイメージ上で統合したものの幅が W_c 以内 (実際は変動を考慮し $1.2W_c$ 以内) に収まるものについて、それが分離文字である可能性が得られる。このような組み合わせによる統合図形に対してもマッチングを行えるように以下のアルゴリズムで新たな切り出し図形集合を C に追加する。

各マップ番号 i に対して

$$w_i + \sum_{j=1}^{n}(b_{i+j-1} + w_{i+j}) \quad < \quad 1.2 \times W_c$$

w_i : 切り出し図形 c_i の横幅 (dot)
b_i : c_i と c_{i+1} の間の空白の幅 (dot)

が $1 \leq n \leq 4$ で成立するならば、図形 c_i に隣の j 個 ($1 \leq j \leq n$) をイメージ上で統合した図形

$$c_i^1, \ldots, c_i^j, \ldots, c_i^n$$

をすべて切り出し図形集合 C の要素に加える。右肩の数字は右側にイメージ統合した図形数を示し、これを仮統合レベルと呼ぶことにする。またこのときマップ番号 i 上での最大仮統合レベルを n と定める (仮統合レベルの最大値は４とする。何故なら分離文字の最大分離数は５である。例「州」)。これに伴い最初の C の要素である $c_0, c_1, \ldots, c_m$ は、仮統合レベル０の図形群であるから、$c_0^0, \ldots, c_m^0$ と書き直すことができる。

結局、すべての統合可能性を考慮した切り出し図形系列として次のような

$$C = \{c_0^0, c_1^0, c_1^1, c_1^2, c_2^0, c_2^1, \ldots, c_m^0\}$$

が出力される。

2.2.3　前処理、特徴ベクトル化部

特徴量には当研究室で開発された方向線素特徴量を用いる。これは文字画像データを 64 × 64dot に正規化後、細線化、さらに、たて、よこ、± 45 ° の 4 方向に線素化して 49 の小領域で数えあげ、4 × 49 = 196 次元のベクトルとするもので、特に活字認識に高い性能を有すことが報告されている。

これに従い、以下のような過程で切り出し図形イメージを特徴ベクトルに変換する。

- ノイズ除去、スムージング
- 正規化 (線形、可変幅)
- 細線化 (12 回)
- 線素化
- 特徴ベクトル化

可変幅正規化の提案

方向線素特徴ベクトルは基本的に線素の数え上げであるから、そのノルムは元図形の構造的複雑さを示すと言える。ところが点や線など構造が単純で幅の小さい図形の場合、一律に通常の大きさの文字と同幅に正規化されることにより巨大な黒画素に拡大され構造的特徴を失ってしまう。同時にこれらは通常の細線化過程では十分に細線化されないため線素が正確に数えれれず、通常の文字よりも著しく大きいノルムをもったベクトルを生じてしまう。またこれらにはノイズ的変動も大きく影響する。これは特に今回のように分離文字の部分パターンという、点、線等単純小図形を多く扱わねばならず、しかもそれらと通常の文字を比較評価する場合には大きな問題となる。このような一律正規化の弊害を取り除くため以下のような可変幅の正規化法を採用した。

通常の正規化幅が $N_w \times N_w$ dot, $N_w = 64$ である場合

x または y 方向について、文字幅 w に対し新しい正規化幅 N_{aw} を各文字毎に設定する

$$N_{aw} = \begin{cases} \frac{1}{4}N_w & (w < W_a) \\ N_w & (w \geq W_b) \end{cases}$$

これに順うと、正規化後の大きさによって、

type 0 :　64 × 64 dot

type 1 :　16 × 64 dot

type 2 :　64 × 16 dot

type 3 :　16 × 16 dot

の 4 種類の正規化タイプが生ずる。

切り出し図形をベクトル化する時には $W_a = W_b = \frac{1}{4}W_c$ (W_c : 切り出し図形群の最大高さ) として正規化を行なう。

一方、辞書ベクトルの作成時にも、同様な文字基本幅を辞書文字の最大幅から W_d と定め、これを用いて

$$W_a = \frac{1}{4}W_d,\ W_b = \frac{3}{16}W_d$$

とする。閾値に幅をもたせることにより、あいまいな大きさの文字については複数の正規化タイプでベクトル化し、入力の変動に対応出来るようにしている。

2.2.4　ベクトルマッチング部

各切り出し図形の特徴ベクトルと辞書ベクトルの間でマッチングを行なう。辞書に登録された字種は、ＪＩＳ第１水準の漢字および仮名、英数字、記号などをふくむ全角文字 3303 字種と半角の英数字 62 字種である。

マッチングはユークリッド２乗距離を用い、距離の小さい順に１～10 位候補までを結果として出力する。

2.2.5　評価・統合部

ここでは、マッチング結果から統合可能性のある部分については統合・非統合の判定を行い認識結果である文字列を出力する。なお、マッチング部より出力されるのはマッチング距離の小さい順に 1～10 位の 10 候補であるが、本章のアルゴリズムで実際に評価されているのは１位のもののみである。

まず切り出し図形の id 情報から元のシーケンスを再構成し、更にこれらを次の２系統に分類する。

系統１： single path 部　マップ上で最大仮統合レベル０が１個以上連続する部分で、この部分では統合の可能性はなく、また情報はマッチング結果のみなので、それぞれの切り出し図形に対する１位候補の文字からなる文字列を結果とする。

系統２： multi path 部　上記以外の部分で統合・非統合の判定を要する部分である。この部分の判定アルゴリズムについて説明する。

統合判定

系統２の部分は各切り出し図形を枝とし、統合の選択による文字列の決定がパスの選択になるようなグラフと考えることが出来る。このような展開法としては候補ラティスと称する方法が報告されている [2]。

各図形の認識結果から、その信頼度にあたる数値を評価スコアとして各図形に与える。図形に対応する枝の重みはこのスコアに統合数 (=仮統合レベル＋１) を重みとして掛けたものとする。評価関数としては全経路の重みの総和を用い、これを最大とするような経路をもって統合を選択することにする。但し経路選択自体は数学的考察範囲であり、実用化の段階で適当な方法を考察するとして今回は特に考慮せず、全可能性を評価して決定している。

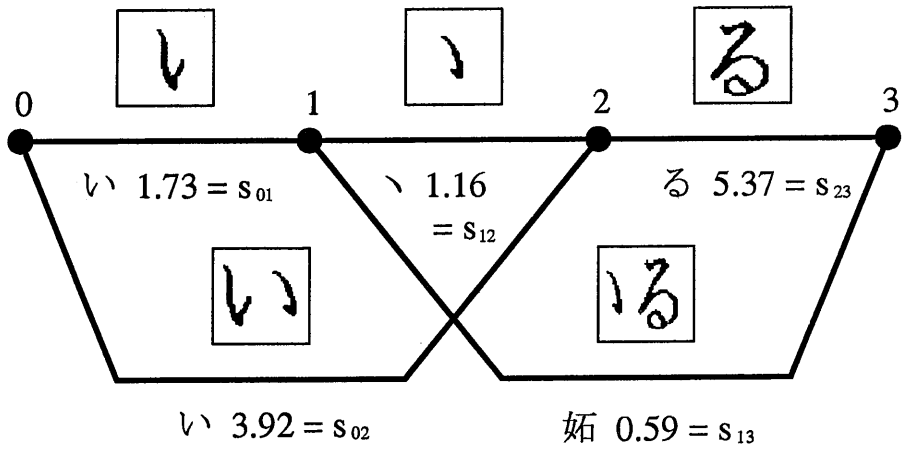


図 2.1: 系統 2 の評価パスグラフと評価関数

上図 2.1は切り出し図形系列からグラフへの展開の例である。箱に囲まれたのが切り出し図形、枝の下の文字と数字はその切り出し図形への 1 位マッチング文字及び評価スコアである。評価スコアは後述の方法で求めるとし、これから各枝の重みが

$$P_{ij} = (j - i) \times s_{ij}$$

s_{ij} ： ij 路図形認識結果の評価スコア
P_{ij} ： ij 路の評価スコア

によって計算される。

そして評価関数

$$F = P_{0k} + P_{kl} + \cdots + P_{mn}$$

が最大となる経路を認識結果とする。ここでは「いる」が最適パスとして選ばれる。

評価スコアの与え方

各図形に対する評価スコアの与え方は、この評価・統合部を作成する上で最も重要であると言える。

信頼度をスコアにするという点からまず、マッチング距離の逆数を用いるという方法が考えられる。すなわち

$$s_{ij} = \frac{1}{d_{ij}}$$

s_{ij} : ij路切り出し図形に対する評価スコア
d_{ij} : ij路 1 位候補のマッチング距離

しかしノルムが小さく構造が単純な図形、特に可変幅正規化によって通常の文字より小さく正規化されたものに同志のマッチング距離は、通常の正規化をなされた文字のマッチング距離が 30〜100 なのに対し、0〜20 と著しく小さい。こういった小図形のマッチング距離に対してはその逆数である評価スコアは不相応に高くなり、その小図形を含むような分離文字は部分パターンに対する (誤った) マッチング結果の評価値が高いためにまったく統合と判定されないという結果を予備実験から得た。こういった例は「い」の右部、「慣」の左の点、「 ”」など非常に多いが、誤認識している辞書側のパターンはいくつかの点図形「丶」「,」などに限られる。図 2.3〜図 2.5は 明朝体 8、12、20 ポイント、ゴシック体 12 ポイントの全角文字サンプル 3303 字種を本書で使用した各書体用辞書でマッチングさせ、正解が 1 位候補に来たものについてそのときのサンプル文字のノルムを横軸に、マッチング距離を縦軸にとったものである。前述の問題とを引き起こす文字はグラフ中では集群から下のほうにはずれた x 軸付近のたかだか十数個であることがわかる。

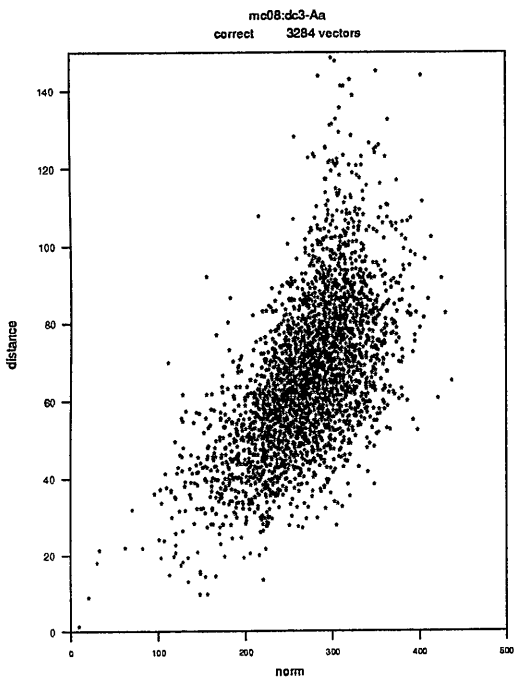


図 2.2: norm-distance (8pt 明朝体)

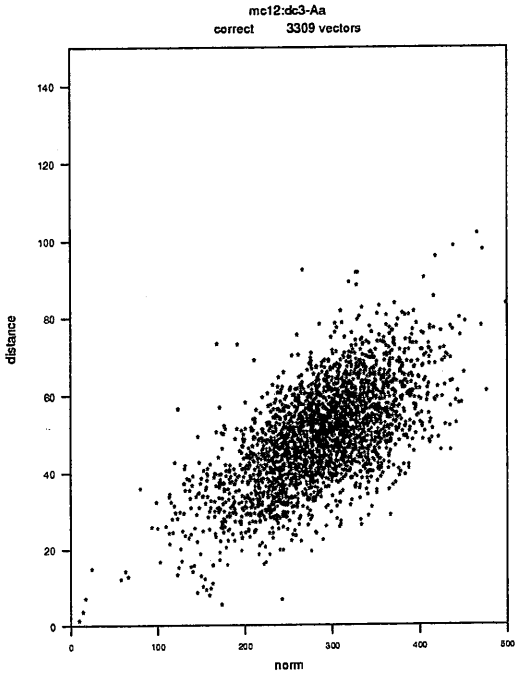


図 2.3: norm-distance (12pt 明朝体)

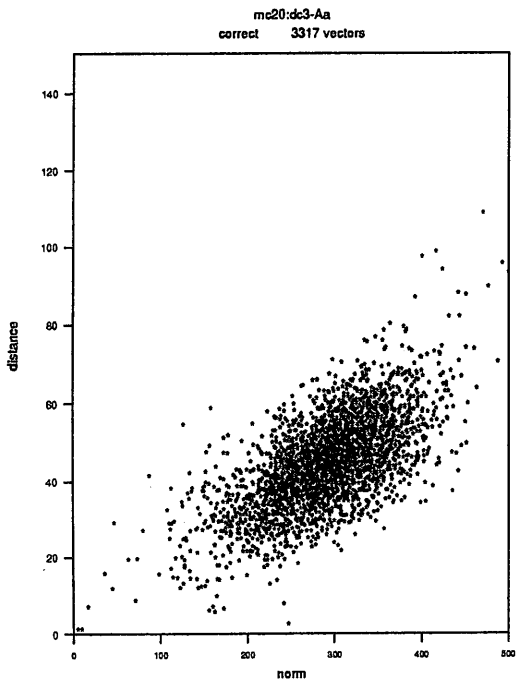


図 2.4: norm-distance (20pt 明朝体)

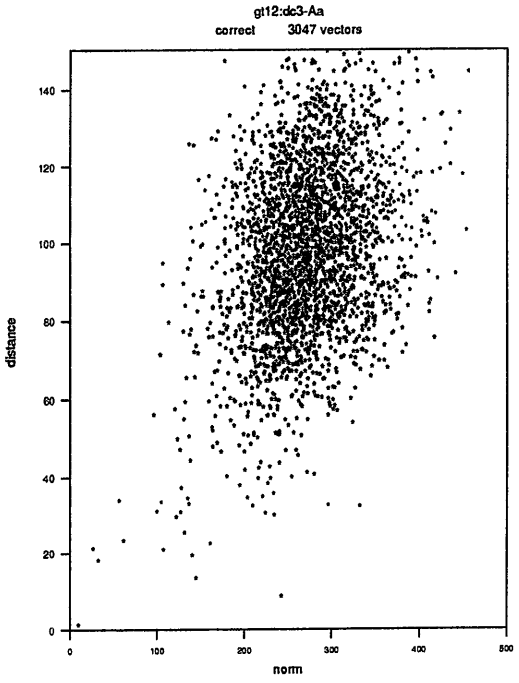


図 2.5: norm-distance (12pt ゴシック体)

以上を考慮した上で以下の３つの評価スコア算出法を考えた。

方法 1

$$s_{ij} = \frac{1}{\max(d_{lim}, d_{ij})}$$

これは、マッチング距離に対する信頼度に限度を設けたもので、ある距離 ($= d_{lim}$) 以下のマッチング距離に対してはもはや有意ではないとして一律 d_{lim} におさえ、過剰評価をもたらすような高いスコアになることを防止するものである。実際この値としては前グラフより２乗距離で 500 と選んだ。($d_{lim} = \sqrt{500} = 22.36$)

方法 2

$$s_{ij} = \frac{n_{ij}}{d_{ij}}$$

s_{ij} : ij路切り出し図形のノルム

方法２ではもっと積極的に、入力ベクトルである候補図形の特徴ベクトルのノルムによって距離値に一種の正規化をほどこし、一律な比較を有効にしようとするものである。ところで、入力ベクトルのノルムをかけるという正規化の妥当性についてであるが、再び図 2.3〜図 2.5より検討する。明朝 12,20 ポイントのように良質のフォントに対してはノルムとマッチング距離が幅はあるものの傾向としてはリニアになっており、妥当性が認められる。しかし、若干のつぶれを含む明朝体 8pt や特に線の太いゴシック体では分散が大きく、故に一般には方向線素特徴量を用いた

マッチングにおいて必ずしも入力ノルムとマッチング距離間にリニアな関係があるとは言えない。そのためノルムの大きな候補図形ベクトル入力に比べノルムの小さな文字が過少評価される危険性も指摘される。しかし図からもわかるとおり少なくとも「有意な 1 位候補に対して小さなノルムの文字が大きなマッチング距離をとることはない」という事実から、過小評価の悪影響は少なく、また方向線素特徴量ではある程度ノルムの大きな文字のほうがマッチングが高精度であることから、むしろこちらを過剰に評価する分には問題は生じないと考えられる。

方法 3

$$s_{ij} = \frac{n_{ij}}{\max(d_{lim}, d_{ij})}$$

2 の正規化をもってしても微少点図形の過剰マッチング評価を抑制することは難しいので 1 の制限を組みあわせたものである。

2.3　認識実験 (1)

本章で作られたシステムを用いて認識実験を行い、統合部の評価スコアの算出方法について考察する。実験に用いた文書データは以下の 3 種で、いずれも日本語 TEX により整形した横書きの文書である。

text1 :　新聞の社説の内容を横書きに直したもので半角文字はまったく含まない。(2278 字)

text2 :　電子ニュースの記事で、半角英数文字を多く含む (2451 字)

text3 :　分離文字と半角文字を主体とするランダム文字列 (2800 字)

このうち **text3** は **text1, 2** の実験のみでは統合の失敗が多くおこらなかったため、わざと間違いやすいように分離文字と半角文字を非常に多く含むように作ったランダムな文字列である。

これら 3 種の **text** について、明朝体では 8, 10, 12pt ゴシック体では 12pt 計 4 種のフォントで合計 12 種のの文書を認識させた。

使用辞書は明朝体テキスト用には明朝体 8, 10, 12, 14, 17, 20pt のフォントからベクトルを作りその平均をとったものを明朝体用辞書に、またゴシック体には 12, 14, 20pt のゴシック体フォントから作成したゴシック体用辞書を用いた。表 2.1はその結果である。

表中で、過統合と統合不十分の覧はいずれも統合判定の失敗を示す。過統合は本来 2 つ以上の文字であるものを 1 つの文字に統合して誤まったケース。統合不十分は逆に 1 つの分離文字を 2 つ以上の文字と判定してしまったケースである。一方最後の覧の統合後誤認識というのは分離文字に対する統合、すなわち文字単位の切り出し位置は合っていたものの認識結果の文字が間違っているケースで例としては「話」→「詰」などのように部首レベルの間違いが多い。一方分離文字の統合判定に関連しないような誤認識については一切考慮していない。

また判定対象文字とは、評価・統合部で系統 2 に属する文字の総数で、総文字数から系統 1 に属した文字数を引いたものに等しい。この中には分離文字の他、半角文字、幅の狭い全角文字なども含まれる。

一般的な文書である **text1, text2** に関してはいずれの方法によっても統合不十分のケースは全統合判定対象文字数に対するミスケースの件数は 1.5%以内に収まっており、過統合による統合判

text	テキスト文字数	統合対象文字数	方法	統合失敗		統合後誤認識
				過統合	統合不十分	
1	9112	2339	1	0	28	1
			2	0	35	1
			3	0	13	1
2	9804	3716	1	0	26	1
			2	0	21	1
			3	0	5	1
3	11200	6858	1	0	333	18
			2	1	129	20
			3	9	79	27
all	30116	12917	1	0	382	21
			2	1	178	23
			3	9	97	30

表 2.1: 認識実験結果 (1)

定ミスはおこっていない。中でも、方法３のスコア法によれば統合失敗件数の率は 0.5%以下となり、一般的なテキストに対してはこの統合手法で精度の高い統合が行えることが示された。

一方分離文字と半角文字が連続することの多い **text3** では方法１では約 5%もの統合不十分の失敗が生じている。このミスの具体例を見ると、大きく３つにわけられる。

1. 「い」「慣」など部分パターンが単純小図形
2. 「 ” 」「∴」など統合前後で小図形
3. 「腕」「信」など部分パターンが有意文字

このうち、ミスのケースのほとんどは１、２のように前節で指摘した点図形の過剰評価によるものであった。単純小図形の過剰評価を防ぐ手法として距離の下限を設ける方法１では「慣」のように微少な点図形に対してはいくらかの効果が見られるものの、１のケースの中でも「い」等のようにノルムのやや大きい点図形には効果はうすく多数のミスが発生した。対してノルムで正規化を行なう方法２では１のケースのミスはかなり少なくなっている。また、ノルムの非常に小さい点図形に対してや２のようなケースは方法１と２の組みあわせである方法３が効果的で、方法１や２のミスケースの多くを正しく統合することが出来た。

方法３で統合出来なかったものの中には３のように部分パターンが通常辞書に登録された通常文字として認識されたため評価値が高くて非統合と判定されたものが多かった。この対策には何らかの形で分離に関する辞書的な情報を導入することが有効ではないかと考えられる。また２のケースによるミスも多い一方で、過統合も９件生じている。これはすぐ隣の句点を誤統合してしまうというパターンで、距離値評価の下限に対する弊害と言える。こうした単純小図形に対する評価をこれ以上パラメトリックな手段で調整するのはむずかしく、限界も存在すると思われる。よってこの件に関しても別手法の組み合わせによる高精度化が期待される。

2.4　まとめ

本章で構成した文書認識システムは方法３の評価スコア算出法を用いることにより良好な統合判定が可能であることがわかった。

現方法で解決出来なかった統合失敗のケースは主に分離後の部分パターンが有意文字に類似している場合である。又もうひとつの誤統合の原因となる単純小図形の過剰評価の問題と点図形の過統合の問題は互いに相反する問題で、現方法のような単純比較による統合法ではパラメトリックな手段での解決は難しい。

そこで、続く第３章ではマッチング時の辞書に分離文字の部分パターンを追加し、この部分パターン辞書を利用した統合の高精度化法について検討する。この方法のねらいは、分離文字の部分パターンを統合図形の一部として認識させることによりその評価値を統合図形の評価値に加算して。パラメトリックな比較への依存度を減らして統合の確実さを増すこと。さらに同時に統合図形の誤認識をも部分パターンによるチェックで減らすということである。

又、同問題への対応法としては、辞書に文字幅の情報を (基本幅に対する比という形で) 登録しておき、判定の際用いるという手法もあり、上のような部分パターンが有意文字という例なら簡単に判定が出来ることが予想される。しかしこのように明な幅比較はオープンな認識には弱く、例のような特殊なケースのみに適用して有効であるとしてもその適用には検討を要す。よって今回の研究では扱わない。

第 3 章

部分パターン辞書の利用

3.1 まえがき

本章では、第２章で構成した文書認識システムの分離文字統合部を高精度化するために部分パターン辞書を導入する。ここでいう部分パターン辞書とは、分離文字の部分パターンイメージを特徴ベクトル化したベクトル辞書であり、前章の認識に用いられた通常辞書に対応するものである。また各文字の分離数などの情報も関連して持ち、統合の情報として利用することができる。

この部分パターン辞書導入の意図としては

- 分離部分パターンの認識結果から統合文字へ加点することによって統合の信頼度を上げる
- 統合パターンに対し、非分離文字と誤認識するのを防止する
- 更に、分離文字の類似文字集合間での誤認識の減少効果

があげられる。

次節では、部分パターン辞書を作成しそれに関する諸問題を考える。次にこれを統合部に導入するためのアルゴリズムを考え、再び認識実験を行なってその有効性を検討する。

3.2 部分パターン辞書の作成 (横分離文字部分パターン辞書)

3.2.1 分離文字の抽出

通常辞書作成用のイメージデータ (全角 3303 字種、半角英数字には分離文字なし) から横分離文字を抽出する。イメージ上で横方向のヒストグラムに切れ目のある文字を (横) 分離文字とする。このイメージはプリンタ出力をイメージスキャナで取り込んだものである。

明朝体およびゴシック体数種のイメージデータ (各１セット 3303 文字) から分離文字を抽出した。各フォントより抽出された分離文字数を文字数順に並べたのが図 3.1である。また使用したフォントの一覧および略記は表 3.1のとおりである。

字体	ポイント数	フォント名（略称）
明朝体	5,6,8,10,12,14,20,25	5mc～25mc
ゴシック体	6,12,25	6gt,12gt,25gt

表 3.1: 分離文字抽出サンプルフォント一覧およびその略称

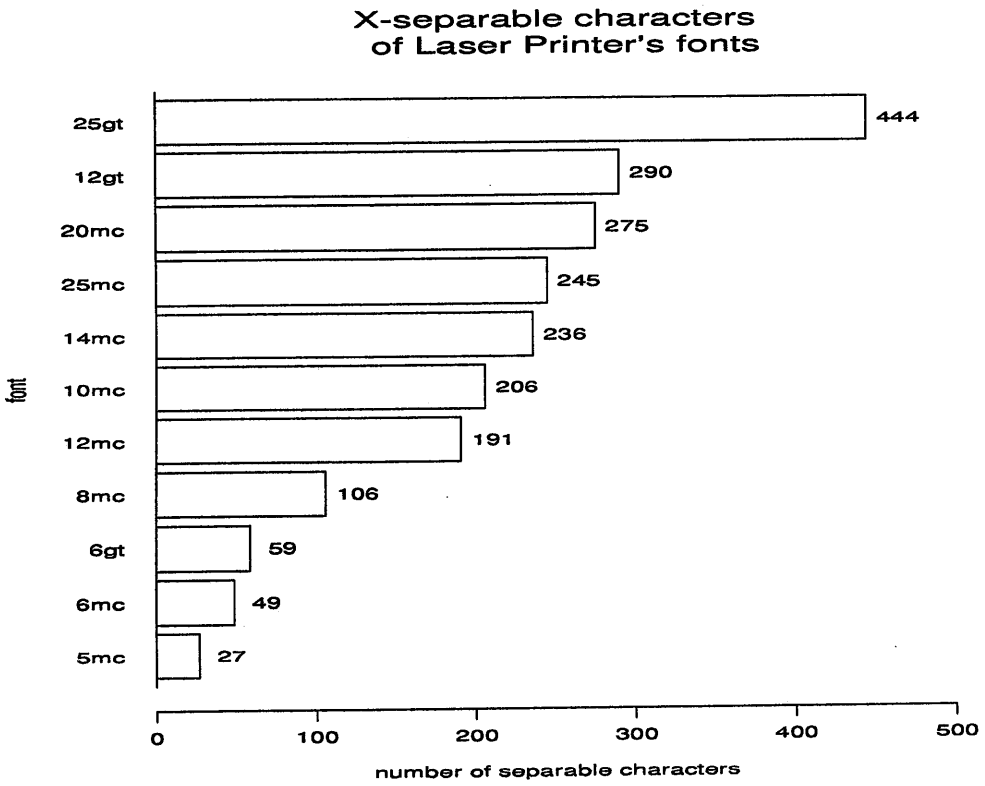


図 3.1: レーザプリンタ文字中の横分離文字数

これを見ると、全角文字中の分離文字数が字体によってことなることは当然予想されたが、同一字体であってもポイントによって大きく異なることがわかる。

一般にはポイントが大きい程そのイメージデータは本来の字体構造に忠実であると考えられる。著しく小さなポイントのフォントでは分離文字数も少ない。これはプリンタ及びスキャナの解像度性能によりイメージ上で「つぶれ」が生じ本来分離する部分が画素で埋まってしまったためである。

しかしこの結果を見ると、分離文字数はポイントの減少に対し必ずしも単調減少ではないことが示されている。文字が小さくなるにつれ分離文字数を増加させる効果の原因は「かすれ」である。「かすれ」と言っても元は長い直線の横棒が２本に分断されるような「かすれ」は当該のように比較的良質な印刷状態のもとではおこらない。実際は、本来の構造で「はらい」や「横棒」が隣の部分と微妙に重なっているためにヒストグラム上で分離がおこらないような文字が、この「は

らい」や「横棒」の鋭った先端がかすれて短くなるために分離するようになる、というケースであることが視察によりわかった。

この「つぶれ」と「かすれ」という一見相反する２つの効果はいずれもポイントの小さなものに対して働く。故に各ポイントでの分離文字数はそのポイントで固有のものとなる。ここで問題となるのは、いくつかのポイントのフォントから分離文字を抽出して、あるポイントでそれが最大であったとしてもその文字集合は他のポイントで分離する文字のすべてを含むわけではないことである。

今回部分パターン辞書を作成する際には、対応する通常辞書の作成に使用されたポイントすべてのセットから分離文字を抽出し、その和集合をもって分離 (可能) 文字集合とすることで部分パターンの登録もれが少なくなるようにした。この部分パターンイメージから部分パターンベクトルを作成、ポイント間で重複するベクトルは平均をとった。また、１つの文字でも分離の仕方が複数あるもの存在する。このような文字について分離数が異なる場合、もしくは分離数が等しくても幅 (基本幅 W_d との比) をチェックして同一パターンとしては疑わしい場合 (分離する場所が異なる場合) はすべて別パターンとして登録する。

3.2.2　部分パターンのベクトル化

ベクトル化の手順は第２章で述べたのと同様である。ここで注意すべきなのは横分離文字の部分パターンは通常文字パターンよりも縦長であり、正規化の際より拡大されるということである。よって正確に線素を数えあげることが出来るためには通常文字では十分とされる回数よりも多くの細線化が必要である。第２章で用いた 12 回という回数はこれを考慮したものである。又部分パターンには単なる点や線といった幅の著しく狭いパターンが多数存在する。このようなパターンには前出の可変幅正規化が効果的である。

3.2.3　部分パターン辞書のクラスタリング

これまでの手順で作成した部分パターンのベクトル辞書には異なる文字から抽出された同形、相似形、あるいはごく類似したパターンの組が多数存在する。このままの辞書で部分パターンのサンプル (12pt の部分パターンベクトル) に対するマッチングのテストを行なったところそのような類似パターンについて正しいマッチングを行なう (マッチング結果で正解が上位候補にくる) のは無理なことがわかった。

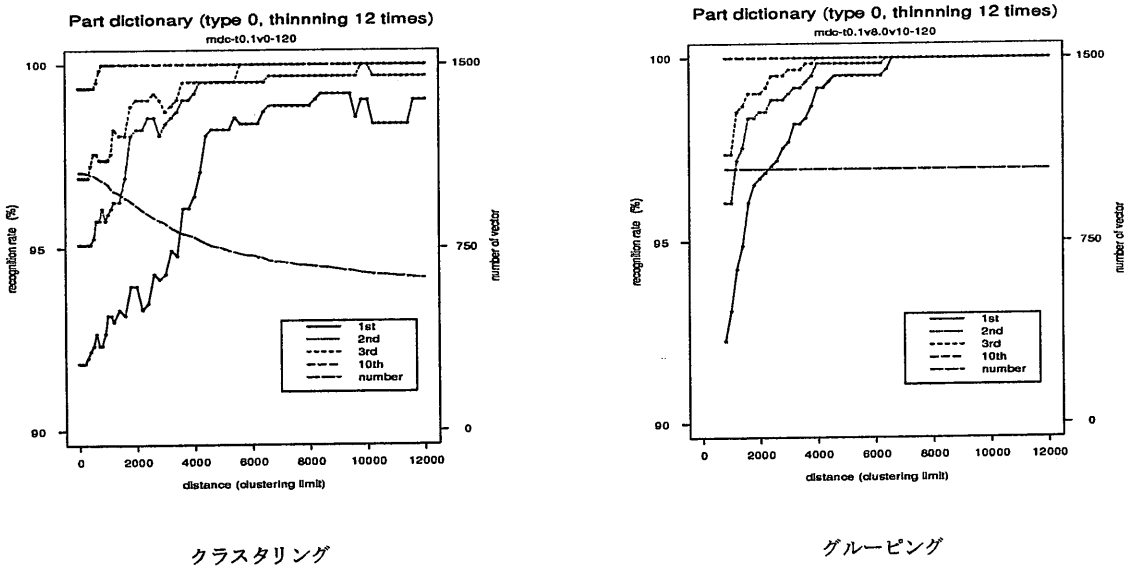
クラスタリング

これら類似パターンのうち、イメージ上では殆ど同形とみなされ、そのベクトル同志が非常に類似している…すなわちベクトル空間上で非常に近接しているものについては、それらをベクトル的に区別するのは不可能である。もっとも、個々の文字認識という観点からはそれらは対をなす部分パターンをもって他の文字と区別されるので、そのような類似パターンは区別する必要もない。この例としては、「は」、「ば」、「ぱ」の左部、「怪」、「快」の左の点などがある。そういったベクトルの組に対してはクラスタリングを行なう。クラスタリング手法には、最小距離 (階層) クラスタリングを用いる。これはある距離値の上限であるクラスター距離を与え、全ベクトル中で距離が最小でかつこの閾値以下であるようなベクトルの対をその平均からなる代表ベクトルで置きかえていく手法である。

グルーピング

イメージ上では類似であるようなものでもベクトル化すると差異があるような組はマッチングの際には誤認識の対象になり、対処が必要である。しかし前述のようなクラスタリングを用い、クラスター距離を大きくとってこれらをクラスタリングしてしまうと元のベクトルと平均化された代表ベクトルの差が大きくなり、逆にマッチング精度が低下するという弊害がある。このような組に対してはベクトル辞書のコピーに対し先と同様のアルゴリズムでこれらがクラスター化されるような大きめのクラスター距離でクラスタリングを行ない、結果的にどのベクトルがどのクラスターに属すかという情報のみをグループ情報として得る。一方ベクトル辞書のほうは全く変更しないでこのグループ情報を付加することにする。これをグルーピングと呼ぶことにする。

実際今回の部分パターン辞書作成にあたっては、統合部への評価導入実験という目的からマッチングテストで正解候補が上位３位に入るような辞書の作成を目指した。もっとも、統合パターンの評価を助けるのに部分パターンのマッチング結果を用いるという分には適度なグルーピングのみで十分であろうとも考えられる。しかしあきらかに同値なベクトル組が存在することはいたずらに辞書数を増加させ速度低下を招くので最小限のクラスタリングは行なうようにした。また、過度のグループ化によっては部分パターンを組みあわせたとき元の文字が一意に定まらなくなるケースが生じる、(例「卵」と「卯」の左右のパターンがそれぞれ同じグループに属すような状態)今回マッチングテストにおいて１位マッチ結果の正解率が 100%を達成するまでグルーピングを行なった場合「は」「ば」「ぱ」の右パターンが同じグループに属してしまうことがわかった。ここでは、そういった類似文字を部分パターンを使って区別したいという意図があるのでこれを目安にし、グループ化されたパターンの構成を視察してグルーピングの限度を決定することにした。



クラスタリング　　グルーピング

図 3.2: 明朝体部分パターン辞書のクラスタリング、グルーピングテスト

図 3.2は明朝体用のベクトル辞書中正規化のタイプが 0 のもの (通常の正規化をおこなったもの) をそれぞれクラスター距離を変えてクラスタリング、グルーピングした状態でのマッチングテスト結果で、横軸にはクラスター距離 (2 乗距離)、縦軸には 1,2,3,10 位の累積認識 (正解) 率と辞書中のベクトル要素数がプロットされている。ここで認識 (正解) とは入力サンプルの部分パターンに対し、マッチング距離の小さい順に求まった候補が入力と同じ部分パターン (元文字、分離数、パート数 (何番目かということ) が一致) であるか、それと同一グループのパターンである場合を指す。

左のグラフよりクラスタリング用のクラスター距離を 10 位認識率が 100%となる 800(2 乗距離) と選んだ。これは統合がマッチング結果の 10 位まで考慮していることと、クラスタリングは最小限で試験するという理由からである。また、右のグラフよりグルーピング用のクラスター距離は 3 位認識率が 100%となる 5800(2 乗距離) と選んだ。正規化タイプ 0 以外のもの 3 つについてはそのノルムが正規化タイプ 0 のものとは大きく異なり同時に扱うことは不適当である。よって正規化タイプ 1 ～ 3 のベクトルについても同様なテストを行いクラスタリング、グルーピング双方のクラスター距離を決定した。ゴシック体辞書についても同様である。具体的数値を次表に挙げる。

フォント	クラスタリング		グルーピング	
	type0	type1～3	type0	type1～3
明朝体用	800	200	5800	2400
ゴシック体用	600	1000	6000	2800

表 3.2: クラスタリング、グルーピング用クラスター距離

このクラスタリング、グルーピングの距離値は今回は試行ということで若干直感的に選んだが、本来は統合部で評価されるマッチング結果候補数などのパラメータと共に実験で厳密に検討すべき値である。更に将来的には部分パターン辞書をすべて全自動で作成するという目標に対し最も難点であることが予想される。

3.2.4　部分パターン辞書の完成

部分パターンのベクトル辞書には元文字のコード、分離数、何パターン目か、またグループに属す場合はどのグループ番号かがコード順に記載されたグループ情報が添えられ、マッチング結果からその部分パターンの詳細を容易に知ることが出来るようになっている。

以下は今回実験に使用した部分パターン辞書 (明朝体用、ゴシック体用) のデータである。

1. 明朝体辞書用部分パターン辞書

   登録分離文字種 395

   グループ情報数 1073

   ベクトル数 891(うち正規化幅が type0 以外のもの 158)

2. ゴシック体辞書用部分パターン辞書

   登録分離文字種 441

グループ情報数 1080

ベクトル数 882(うち正規化幅が type0 以外のもの 46)

3.3　評価部への導入

部分パターン辞書はベクトルマッチングの際通常辞書と同時に用いる。よってマッチング結果である 1 〜10 位の出力候補中には通常文字と部分パターンが混合して現れる。

3.3.1　切り出し図形スコア評価部の改良

第 2 章の評価・統合部に部分パターンの評価スコアを導入する。

系統 1 の部分は変更はない。系統 2 の部分について、仮統合レベルが 1 以上の図形、すなわち 2 個以上の切り出し図形に分離出来る図形を統合図形と呼ぶことにする。この統合図形 1 つに対してその統合レベルに応じて下のように複数の分離の仕方がある。

level 1 :	ab	→ a+b	1 通り
level 2 :	abc	→ a+b+c, ab+c, a+bc	3 通り
level 3 :	abcd	→ a+b+c+d, ...	7 通り
level 4 :	abcde	→ a+b+c+d+e, ...	15 通り

ある、仮統合レベル n の統合図形に対する分離の仕方のひとつに注目する (このときの分離数を m とする) と、評価対象として

- 統合図形に対するマッチング結果： 1 位〜10 位
- 部分パターンに対するマッチング結果： 1 位〜10 位 × m

が使用できる。各マッチング結果は第 2 章で述べたスコア評価法を用いて評価されるとする。

統合図形 j 位のマッチング結果が通常文字 x であるとき、各部分パターン毎のマッチング結果から統合図形 x の部分パターンであることに矛盾しないもの (分離数、パート数) のうち最も評価の高いものを抽出する。この合計 0 〜m 個の部分パターンの集合を S_p とする。これら部分パターン集合の評価スコアの総和を統合図形 x に対する加点値 a として x の評価スコアに加点操作する。

分離の仕方が複数ある場合対応する他の部分パターンの組み合わせに対しても同様の操作を行い、得られた加点値を加点操作した評価スコアが以前のものより大きいならばそれを新たな評価スコアとする。

この操作を統合図形の 1 〜10 位マッチング結果に対し行なうことにより、部分パターンの評価値を考慮した新たな評価スコア順の 1 〜10 位候補を得ることが出来る。

3.3.2　評価スコアへの加点方法

ij 路の統合図形の評価スコア s_{ij} への加点方法

加点モード 1

$$s_{ij} = b_{ij} + a$$

加点モード 2

$$s_{ij} = (b_{ij} + a) \times \frac{k}{m}$$

b_{ij} : 統合図形の認識結果に対する評価スコア
a : 加点値 $\sum_{s \in S_p} s$,
m : 分離数
k : S_p の要素数、 $0 \leq k \leq m$

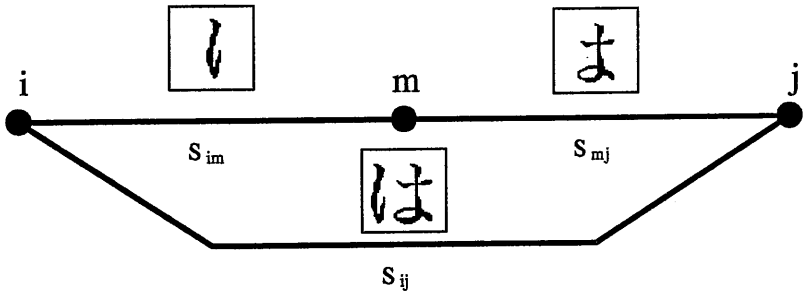
加点モード 1 は単純に加算する方法で、この方法によれば各部分パターンのマッチング結果で統合図形の部分パターンと認識され、その評価値が誤認である通常文字のマッチング結果の評価値よりも高いか、少なくとも同等の値であればパス選択は確実に統合側に有利になされる筈である。

加点モード 2 は統合図形のあるマッチング結果に対しその部分パターンが矛盾なく認識されなければ統合図形のマッチ結果自身の評価値をも減点しようというものである。もし統合図形のマッチング結果が分離文字でなかったならその結果は 0 点となり候補からははずされる。これはすなわち過統合や統合誤認識をより積極的に排除しようというものである。

次頁に仮統合レベルが 1 の場合のスコア加点の例を示す (図 3.3)。認識結果は 3 位まで考慮するとし、各図形に対する評価スコアとともに示した。この中で「は-2/2」というのはマッチング結果が分離数 2 の元文字「は」の 2 番目の部分パターンであるということをあらわす。また、「は-1/2g] の g は「は」の 1 番目の部分パターンで代表されるグループにマッチしたという意味で、このグループには「は」「ば」「ぱ」の 1 番目の部分パターンが登録されているためそのいずれの文字の部分パターンとしても等しく評価される。

例：仮統合レベル 1 の場合

図 3.3: 加点の例 (仮統合レベル 1 の場合)



「は」の認識結果		「い」の認識結果		「ま」の認識結果	
は	2.63	‘	1.85	は-2/2	3.32
ば	2.02	は-1/2g	1.80	ま	1.16
ぱ	1.86	旧-1/2	1.12	ぱ-2/2	1.14

「は」の加点値　$a = 1.80 + 3.32$　, $k = 2$
「ば」の加点値　$a = 1.80 + 0$　, $k = 1$
「ぱ」の加点値　$a = 1.80 + 1.14$　, $k = 2$

評価スコア s_{ij}

加点モード 1

は　$s_{ij} = 2.63 + 1.80 + 3.32 = 7.75$
ば　$s_{ij} = 2.02 + 1.80 + 0 = 3.82$
ぱ　$s_{ij} = 1.86 + 1.80 + 1.14 = 5.10$

加点モード 2

は　$s_{ij} = (2.63 + 1.80 + 3.32) \times \frac{2}{2} = 7.75$
ば　$s_{ij} = (2.02 + 1.80 + 0) \times \frac{1}{2} = 1.91$
ぱ　$s_{ij} = (1.86 + 1.80 + 1.14) \times \frac{2}{2} = 5.10$

3.4　認識実験 (2)

評価スコアの算出には第２章で効果のあったノルムで正規化した距離の逆数を用い、更にそれぞれ距離の下限値を用いるかどうか２種類ずつ行なった。以下がその結果である。

text	文字数	加点モード	過統合		統合不十分		誤認識	
			d_{lim} 無	d_{lim} 有	d_{lim} 無	d_{lim} 有	d_{lim} 無	d_{lim} 有
1	9112	なし	0	0	35	13	1	1
		mode 1	1	1	0	0	2	1
		mode 2	1	0	15	15	3	2
2	9804	なし	0	0	22	5	1	1
		mode 1	0	0	0	0	1	1
		mode 2	0	0	6	4	5	5
3	14000	なし	1	9	121	79	20	27
		mode 1	2	2	17	4	14	10
		mode 2	3	2	24	11	18	16
all	30116	なし	1	9	178	97	22	29
		mode 1	3	3	17	4	17	12
		mode 2	4	2	45	30	26	23

表 3.3: 認識実験結果 (2)

過統合、統合不十分、誤認識の各欄は表2.1のものと同じ意味である。またそれぞれ評価スコア算出時に距離の下限を用いないときと用いたときについて並べてある。加点モードについて、「なし」の場合はそれぞれ d_{lim} のあるなしが第２章の認識結果表中の方法２、方法３に相当する。mode1 の覧は単純加算の加点モード１の結果、mode2 は部分パターンのマッチ率をかけた加点モード２の結果である。

加点モード１を用いた場合

text1 および **text2** では統合不十分に関する失敗を全くなくすことが出来た。これは距離のリミットを用いないスコア算出法においても達成出来ており、評価値のノルム正規化と組み合わせることで単純小図形の過剰評価に起因する統合不十分をノンパラメトリックに改善し高精度の統合が可能になることを示している。トータルで見ても統合不十分のケースは全統合対象文字に対し加点なしの場合からの改善度は

$$0.98\% \rightarrow 0.09\% \quad (d_{lim}\text{なし})$$
$$0.53\% \rightarrow 0.02\% \quad (d_{lim} = \sqrt{500})$$

と非常によい成績と言える。またこの段階での統合不十分のミスケースはすべて「 ”」「“ 」に対するミスであり、第２章の方法では困難とした、分離後の部分パターンが有意文字であるようなケースはすべて統合は正しく行われた。また統合不十分が大幅に改善されたにもかかわらず、統合後の誤認識も減っている。過統合も d_{lim} なしで２件増えているのみで、d_{lim} 有でも加点なしの場合のように増加しておらず、過剰統合も防止されていると言えよう。この過統合の具体例は１例で「 ; }」→「 ”」であった。

加点モード 2 を用いた場合

主に過統合と統合後の誤認識を防止するための加点法であったのだが、もともと過統合の例が少ないこともあり、加点モード 1 に比べ良いとは言えず。誤認識も多い。統合不十分も加点モード 1 よりずっと多く、良い方法とは言えないようである。

この方法がうまくいかなかった原因としては、統合図形のマッチング結果に比べ部分パターンのマッチングの精度が低く、正しい統合候補に対し減点に働くことが多いということが考えられる。こういった辞書の性能の問題に関しては前節の部分パターン辞書作成の部分にたちもどる必要があるが、辞書の作りかたによっては良い結果が得られる可能性もあり、検討すべき課題である。

3.5　まとめ

本章では、分離文字の部分パターンベクトルを登録した部分パターン辞書を作成し、通常辞書と合成してマッチングに利用した。これにより得られた部分パターンとしてのマッチング結果をその統合、分離文字のマッチング結果評価値に単純加点することで前章で作成した分離文字の統合部を大幅に高精度化することが出来た。

この方法をもっても統合不十分の失敗となるような「”」→「,,」「∴」→「…」という間違いは現方法での点図形に対するマッチングの精度限界に起因する。こうした点図形を含む線等の単純小図形に対しそのわずかな特徴をとらえて正しく認識するのは現行の方向線素特徴量を用いた方法ではむずかしい。とくに正規化幅を$\frac{1}{4}$と小さくしているのベクトル化した時点で個々の違いを十分に反映出来ていないとも言える。しかし正規化幅を大きくとって細線化回数をさらに増やしたとしてもノイズなどの変動も大きくなりかえってマッチング精度が落ちてしまうことがわかっている。このため単純小図形の構造的特徴に着目するような別の特徴量の導入も有効と思われる。

部分パターン辞書導入で問題となるのは処理速度の低下であろう。現在辞書数は部分パターンによって約 1000 文字増加しベクトルマッチングに要す時間は 1.3 倍になる。一方部分パターンの評価 (加点) に要する時間は全くのテキスト依存である。ここで時間を要すのは部分パターン辞書のグループ情報からの索引であり、その回数は仮統合レベルの大きさによって指数的に増大する。しかしそういった増大を引きおこすような分離数４、５の文字が通常の文書で連続することは少なく、評価に要する時間はさほど重視する必要はないと言える。

また今回の実験では統合図形、部分パターン双方とも１０位候補までチェックをしたが、特に部分パターンについては辞書ベクトルが効果的にクラスタリング、グルーピングされていれば統合の精度を下げずにある程度下位候補を切り捨てることも可能である。適度なクラスタリングはまた辞書登録ベクトルを減らすことになり、マッチング時間を減少させること出来るので統合精度を考慮しつつパラメータを設定するような検討が必要である。しかしこういった辞書作成のパラメータ調整は自動化を目指す必要があるためたいへんむずかしい問題である。

部分パターンのマッチング精度が上がれば統合後の誤認識を減らす方の効果も期待でき、部分パターン辞書導入の意義がさらに増すと言えよう。

入力文書中で部分パターンに登録がないような分離をした文字は統合図形に対する素点のみで評価されることになる。辞書作成時に印刷状態を変えてみるなどの工夫によって出来るだけ多くの分離可能文字を登録することが検討されるが、すべてに完全に対応出来る辞書を作ることはむずかしく、登録漏れに対してもマイナスにならないような評価法を考える必要がある。今回の認識実験でもわずかであるがこのケースも存在したが、単純加点法のためマイナス効果となることはなくそのためのミスは生じなかった。

第 4 章

結論

4.1 本研究のまとめ

第２章では方向線素特徴量を用いた文字認識システムで通常辞書のマッチング結果のみから分離文字を統合するシステムを検討した。マッチング結果に対する評価値にノルムで正規化した距離の逆数を用いて単純比較によってもかなり正確に統合判断がおこなえることを示した。統合ミスの内容としては部分パターンが単純小図形であるケースと部分パターンが有意文字であるケースがほとんどであり、その中でも単純小図形の過剰評価に伴うミスを防ぐ方法としてマッチング距離の評価値に対し限度を設ける手法が単純だが効果的であることがわかった。

第３章では部分パターン辞書を作成しそのマッチング結果を評価・統合部で統合図形の評価値の加点値として考慮することで第２章の方法で統合できなかった部分パターンが有意文字であるケースのすべてを正しく統合することができた。また部分パターンが単純小図形であるケースにも効果があり、一般的に高精度の統合が可能であることを示した。

4.2 今後の課題

第２章で作成した文書認識システムは今回認識実験対象として文字サイズ一定の横書き文書のみを扱ったが、研究の核である統合部は現在のところベクトルマッチングの結果のみを評価しており、この部分では対象を固定していない。そのため、切り出し部および辞書を修正すことによりサイズ可変の文書、あるいは手書き文書も文字切り出しにも有効であると思われる。しかし、実験段階から実用レベルを考えるにあたっては指針をはっきりさせる必要がある。例えば今回の例のように書式を限定するなら、文字幅のような情報を積極的に利用し、通常文字もあわせてトータルで認識率100%の文書認識システムを目指すことである。また、システム全体で応用度を広くもとうとすると問題となるのが第２章の切り出し図形統合部でおこなった仮定である。この仮定は少なくとも認識対象と辞書間で文字スケール(縦横比等)が同じであることを期待している。例えば現存のシステムは簡単な変更でそのまま縦書きの文書認識にも対応することが出来るが、認識対象を新聞文書とした場合新聞文字のフォントの縦横比のちがいにより切り出し部のアルゴリズムに問題が生じる。こういった認識対象の変動を許容し、さらに無駄のない切り出し図形生成アルゴリズムというものを検討しなければならない。

第３章の部分パターンの利用についてはかなりのメリットは認められたものの実行速度や使用メモリなどのデメリットも存在し、有効性をより認められるような検討が必要である。このための一番の課題は辞書の構成法である。本文中で述べた分離可能文字の抽出法、クラスタリング、グ

ルーピングの仕方についてさらなる実験を通し最適となる手法をみつけること、特に今回のように印刷精度の高い文書に対してのみではなく、より低品質な雑誌など一般の文書に対しても普遍的に有効に働くような部分パターン辞書を構成する方法を見いだせるかどうかが、実用レベルでの目標である「辞書の自動作成」への鍵となるであろう。また、認識システムの基幹としての方向線素特徴量の使用は変わらないとしても、こうした部分パターンの認識に対する同特徴量の弱さが指摘された以上、部分的な特徴量の改良もしくは他の特徴量の補助的使用が検討される。

また文字切り出しを完全に行なうためには接触文字の分離処理が必要であり、これは大きな課題である。

謝辞

本研究を進めるにあたり、全般的な御指導を賜わりました東北大学工学部阿曽弘具教授に心より感謝致します。

また東北大学工学部阿曽研究室の大町真一郎氏、後藤英昭氏、池田啓明氏、越後和徳氏には文字認識の専門分野においてさまざまな御助言、御指導、御協力頂きました。ここに深く感謝いたします。

日々の研究においては計算機環境をととのえていただきました同研究室の成富敬氏、沼田一成氏、中井満氏、福田大氏に厚く御礼申しあげます。

最後に日頃よりお世話になりました同研究室の皆様に心より感謝致します。

参考文献

[1] 孫寧、田原透、阿曽弘具、木村正行：「方向線素特徴量を用いた高精度文字認識」
電子情報通信学会論文誌 (D-II), Vol.J74-D-II, No.3, pp. 330–339 (1991 年 3 月)

[2] 村瀬洋、若原徹、梅田三千雄：「候補文字ラティス法による枠無し筆記文字列のオンライン認識」
電子情報通信学会論文誌 (D), Vol.J68-D, No.3, pp. 765–772 (1985 年 4 月)

[3] 佐藤道弘、木田博巳：「不定ピッチ文字列を含む印刷文書における文字切り出し法」
電子情報通信学会技術研究報告, Vol.88, PRU 88-159, pp. 97-104 (1989 年 3 月)

[4] 越後和徳：「文字認識アルゴリズムの高精度化に関する研究」
東北大学大学院工学研究科情報工学専攻 修士学位論文 (1993 年 2 月)

[5] 長尾真：「パターン情報処理」
コロナ社 (1983 年 3 月)