修士学位論文

文字特徴量の分布形状を考慮した文字認識用辞書の構成法に関する研究

東北大学大学院 情報科学研究科 情報基礎科学専攻

丸岡研究室

孫 方

平成8年2月15日

目 次

第 1 章

序論

1.1 本研究の背景と目的

情報処理技術の急速な発展にともない、計算機の扱う情報の量はますます増大し、多用化、複雑化している。これまでの通信・情報等の技術の大幅な進展により、手順の決まった計算や処理であれば膨大な量をも短い時間でやりとげることができるようになった。しかし、計算機の発展の速度に比べ、人間と計算機の間の情報交換手段、いわゆるマン・マシンインターフェースはそれほど改善されていない。もし、我々が普段情報交換のために用いている、紙に書かれた、あるいは印刷された文字を自動的に計算機に入力できれば、それは人間にとって訓練の必要がなく、わかりやすく、容易な手段と言える。さらに、近年ではOA化により、ワープロやファクシミリ、コピー機などの事務機器が産業から家庭に広がっており、大量の文書が出回っている。このような大量の文書の保存、変更、管理を計算機によって行なおうという社会的要求が強まってきた。急速に発展している情報化社会のなかで、文字認識技術の重要さは音声認識・画像処理技術と並んでますます重みが増していくと考えられる。

最近の信号処理技術・エレクトロニクスなどの進歩は見る・聞く・読むといった人間の知的活動機能の代行をも可能としてきている。文字認識についての研究は約三十年前から行なわれ、種々の成果が得られている。現在、特定の分野においてはOCR (Optical Character Reader) が実用化されている。そして、社会活動における膨大な事務処理、たとえば伝票

の読みとりなど、科学技術論文や特許情報等のデータベース作成のための入力装置、病院におけるカルテの計算機処理用入力など、広い範囲にわたって使用されるようになってきた。また、これらの装置で読まれたデータは通信回線を通して遠隔地に送られ、蓄積・処理・検索利用されるようになろう。こうして通信・情報技術の進歩に支えられて進展著しい文字認識技術は多彩な情報文化社会形成の重要な一翼を担おうとしている。しかし、現状では認識率 100%の文書認識システムは完成されておらず、どうしても誤認識を生じてしまう。誤認識がある場合には、修正に手間がかかり過ぎるために多くの労力が必要とされる。このため、文書認識の完全な自動化に大変期待が寄せられている。

本論文では、文字特徴量の分布形状を考慮した文字認識用辞書の構成法について検討する。まず統計的な手法を用い、文字認識で用いられる特徴量の空間上での分布状況を明らかにする。その上で、マルチフォント印刷文字および手書き文字認識に有効な認識手法および認識辞書構成法を提案し、実験によりその有効性を確認する。

マルチフォント印刷文字認識

文字認識において、単一フォントの活字を扱う場合と比較して、マルチフォントの活字文字を扱った場合の認識率はかなり低い。対策として、パターンの変形に強い特徴量の開発 [3]、入力パターン [4] や辞書パターン [5] を変形する手法、字種ごとに複数の辞書を用意する方法 (マルチテンプレート法)[6][7][8] 等が検討されている。

これらのうちマルチテンプレート法はアルゴリズムが簡単であり、既存の認識システムに組み込むことが容易であるという利点があるため、これまでさまざまな研究が行われてきた。塩野 [6] は、単純類似度法の辞書を複数用意する手法を用い、手書き文字を対象に大規模な実験を行ってその有効性を確認している。八代 [7] は、手書き文字を対象にクラスタリングアルゴリズムと一字種あたりのテンプレート数を様々に変えて実験を行っている。クラスタリング法は 3 種類の中で LBG 法が一番よい、テンプレート数は一字種あたり 32 個のとき一番いい認識率が得られた、などの結果が得られている。加藤ら [8] は、明朝体・ゴシック体・その他の辞書を持ち入れ換えて用いる手法を提案している。エラー率関数を用い、最初に一部を認識し、人が介在して辞書の配置を換える (二次記憶上の辞書のうち、必要なものを主記憶に置いて使う) ことにより、少ないテンプレート数で認識率を向上する

ことを可能にしている。

しかし、これらの研究では同一字種のサンプルパターンの分布 (カテゴリ内分布) のみ考慮してクラスタリングを行っており、異なる字種のサンプルパターン間の分布 (カテゴリ間分布) はほとんど考慮されていない。また、各字種ごとの最適なテンプレート数の判断法についてほとんど触れていない。分布形状の特徴を考慮しない単純なマルチテンプレート化は、認識用辞書が必要以上に大きくなり、記憶容量・計算時間が増加することにつながる。また、同一字種のテンプレートを複数持つことは異なる字種のテンプレート間の距離が近くなることを意味し、逆に誤認識となる字種も現れる等、悪影響も見られる。本研究では、カテゴリ内分布とカテゴリ間分布の両方を考慮することによりこれらの問題点を解決する活字文字認識用のマルチテンプレート辞書作成法として、字種適応クラスタリング法を提案する。まず統計的な手法を用い、文字認識で用いる特徴量の空間上での分布状況を調べる。その結果から誤認識が生じ得る状況を推定し、それを踏まえて誤認識の可能性のある字種のテンプレート数を増やすという操作を繰り返すことで、不必要な複数化を防ぎ、かつ必要なものはさらにテンプレートを増やすという操作が可能になる。字種適応クラスタリング法は、この考え方を基本とし、従来法 (LBG 法) と比較して認識率を下げずに総カテゴリ数を抑えたマルチテンプレートの文字認識用辞書を人が介在せずに作成するものである。

そして、マルチフォントの印刷文字を対象とし、提案した手法を用いて実際に認識用の辞書を作成し、認識実験を行うことで手法の有効性を確認する。

手書き文字認識

手書き文字認識に関しては、電総研提供の手書き文字データベース ETL9B [9] を用いた研究が活発に行われている [10][11]。近年、孫らは改良型方向線素特徴量 [12] を用い、マハラノビス距離を改良した距離尺度を用いて平均 98.24%の認識率を得ている [10]。また、若林らは ETL9B のサンプル数の少なさを非線形正規化を利用して学習サンプル数を増加させることで補い、特徴量の高次元化・特徴選択等により 99.05% の高い認識率を得ている [11]。しかし、これらの研究では 1 位認識率を向上させることのみを目的としている。すなわち、未知入力パターンが与えられたとき認識対象字種の中で最も確からしいと思われる

ものを候補としたとき、それがどの程度正しいのかが認識系の良さの尺度となっていた。

しかし、ETL9B の中にはどのような認識システムを用いても認識不可能と言われている文字が約 700 文字存在する [11]。また、ETL9B に限らず実際に人間が書いた文字には何と読むべきか判別の難しい文字が少なからず存在する。また、JIS 第 1 水準の文字のみを扱う認識系での第 2 水準の文字、誤記等により実際には存在しない文字が書かれている場合、および文書認識時に切り出しミスにより分離文字の各部分が認識対象となった場合など、候補を出力すべきではない場合も多い。どのような場合にも必ず 1 個の候補を出力する認識系では、以上の場合に好ましくない候補を出力することになる。したがって、実用的な認識システムにはリジェクト機能を付加することが不可欠である。一方、個々の文字認識の結果の正しさを保証しようという研究もなされている。阿曽らはパターン整合法のアルゴリズムを定式化し、領域半径の概念を導入し、文字認識において選出された候補の正しさの保証に関する検討を行っている [2]。

本研究では手書き文字を対象とし、認識結果の保証を組み込んだ候補選出アルゴリズムを提案する。具体的には、まず、サンプルパターンにより各字種の文字パターンの特徴量が分布する領域 (特徴領域) を統計的に定める。そして、領域半径の概念を用いて未知パターンが属している特徴領域を推定し、候補選出を行う。さらに、得られた候補の信頼性を検証する。

領域半径の概念を用いる場合にはパターンの分布を正確に表す評価関数が必要となる。マハラノビス距離は適当な評価関数の一つであると思われるが、計算時間がかかり、サンプルが少ない場合に値が正確に求まらないなどの問題がある。マハラノビス距離の計算時間の削減を目的とし、これまで、擬似マハラノビス距離 [13]、改良型マハラノビス距離 [10] などの距離尺度が提案されている。これらは、マハラノビス距離の式のうち低次元の項のみを用いて計算を行うことで計算時間を削減している。しかし、これらの式はマハラノビス距離を近似することを目的としていないため、実際のマハラノビス距離との誤差については考慮されていない。本研究では、マハラノビス距離の問題点を解決し、少ない次元数でよりマハラノビス距離を近似でき、確率的に同等な距離尺度である簡素化マハラノビス距離 (Simplified Mahalanobis Distance) を提案する。そして、真のマハラノビス距離との誤差をシミュレーションにより解析する。また、高次成分を用いずに計算量を削減し、サ

ンプル数が少ないことによる影響を補正する識別器としては修正 2 次識別関数 (Modified Quadratic Discriminant Function; MQDF) が知られている。しかし、本研究では領域半径の概念を用いてパターンがあるカテゴリの特徴領域に属しているかどうかを判断することを目的としているため、2 次識別関数よりもマハラノビス距離を用いるのが適当であると思われる。

本研究のもう一つの特徴は、一次元特徴領域と多次元特徴領域の二種類の特徴領域を用いている点である。具体的には、学習サンプルから一次元と多次元の二種類の特徴領域を統計的に推定し、これらの特徴領域を同時に参照することで、未知パターンの候補を選出し、その信頼性を検証する。分布形状を表すパラメータを正確に推定することができれば、多次元特徴領域は文字認識に非常に有効である。しかし、一般に学習サンプルから推定したパラメータには誤差が含まれ、特に学習サンプルが少ない場合においてその誤差がかなり大きく、悪影響がある。一方、一次元特徴領域の場合は、認識能力は多次元特徴領域と比べて落ちるが、学習サンプルが少ないことによる影響が少なく一次元空間上の分布に限れば正しく表すことができる。本研究で提案する手法は、候補を選出する際に特徴の異なる二種類の特徴領域を同時に参照することで、それぞれの優れた利点を発揮し、より信頼性の高い候補を得ることを可能にするものである。

そして、ETL9B を用い、提案した手法を用いて実際に認識実験を行うことで手法の有効性を確認する。

1.2 本論文の構成

第 1 章 序論であり、本研究の背景と目的について述べる。

第 2 章 文字認識アルゴリズムについて説明する。

第 3 章 文字特徴量の分布を調べる。具体的には、主成分分析法を用いて、カテゴリ間分布とカテゴリ内分布を調べ、領域半径の概念を用いて特徴領域の重なり具合を調べる。

第 4 章 第 3 章の結果を踏まえ、マルチフォント印刷文字を対象とし、文字特徴量の分布形状を考慮したマルチテンプレート辞書の構成法を提案する。そして、本手法を用いて

認識用辞書を作成し、認識実験より有効性を確認する。

第 5 章 手書き文字を対象とし、文字特徴量が分布する領域を推定することで文字を高精度に認識する手法を提案する。そして、手書き文字データベース ETL9B を用いた認識実験より有効性を確認する。

第 6 章 結論であり、まとめと今後の課題について述べる。

第 2 章

文字認識

文字認識とは、入力された画像が何という文字であるかを判断する処理である。文字認識の手法は、大きく次の二つに分けられる。

1. パターンマッチング法

2. 構造解析法

パターンマッチング法では、予め読もうとする文字の一つ一つにテンプレート (template: 原形) を用意しておく。それに対し、入力された未知の文字パターンを重ね合わせ、最も近いテンプレートの文字がその入力文字であると判断する。すなわち、パターンマッチング法はパターン同士の重なり具合で評価し、認識を行なうものである。このため文字の多少の変形やノイズに強く、計算機上で容易に高速に実現できるが、類似文字の識別は難しい。普通は、計算量削減のため、文字画像そのものではなく文字画像から得られる特徴量 (特徴ベクトル) を用いて認識を行う。

構造解析法は文字がどのように構成されているかを解析して総合的な判断を下す。具体的には、線分の接続関係や位置関係などの文字構造に着目し、構造の類似性で認識を行なう。この手法は類似文字の多い漢字や、手書きなどによる変形が大きい文字の認識に有効であるが、特徴量の定義や抽出が難しく、また計算機上での処理に時間がかかる等、問題も多い。

本研究では、パターンマッチング法を用いた。図 2.1に文字認識のアルゴリズムを示す。

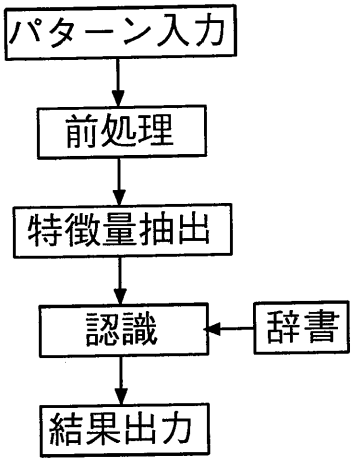


図 2.1: 文字認識のアルゴリズム

2.1　文字入力

文字入力とは、新聞・本・雑誌などの文書を二値画像として取り込む処理である。画像データはイメージスキャナによって入力される。入力された画像データは切り出しの処理により各文字ごとに切り出される。

2.2　前処理

文字を認識する場合の前処理とは、切り出されたイメージデータから特徴量を抽出するために必要なイメージ処理である。ここでは、前処理は、入力文字の線の輪郭を滑らかにするためのスムージング・ノイズ除去、文字の大きさの正規化、細線化と線素化の 4 つの処理によって構成されている。図 2.2に前処理の流れを示す。

2.3　特徴量抽出

パターンマッチング法では、処理の高速化、パターン分離の効率化などのために、文字パターンを特徴量という数値ベクトルに変換する。この過程を特徴量抽出という。本研究では、方向線素特徴量 [1] および改良型方向線素特徴量 [12] を用いた。それぞれの特徴量

について簡単に述べる。

- 方向線素特徴量：

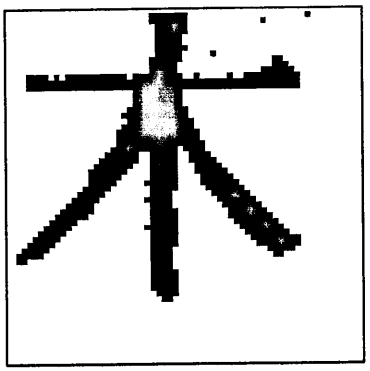
  方向線素特徴量の抽出法は図 2.3のように、まず、64 × 64 ドットの線素化画像の縦横を 8 ドット間隔に分割する。次に、16 × 16 ドットの領域を 49 個定義する (左上を 0 とし、8 ドットずつずらし、左から右、上から下へ順に並ぶ)。各領域ごとに縦、横、斜め 45 °、斜め 135 ° の 4 種類の線素の数を、重みつきでカウントし、4 次元のベクトルとする。1 領域内の重みは図 2.3のようになっている。よって 1 文字あたり 196(= 49 × 7) 次元のベクトルとなる。

- 改良型方向線素特徴量：

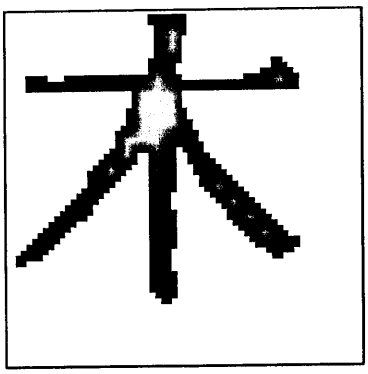
  改良型方向線素特徴量は、方向線素特徴量に対し、以下のような変更を行ったものであり、手書き文字認識により適した特徴量である。

  1. 輪郭線抽出の導入
  2. 正規化アルゴリズム
  3. 線素化アルゴリズム
  4. 外側加重による線素抽出法の導入

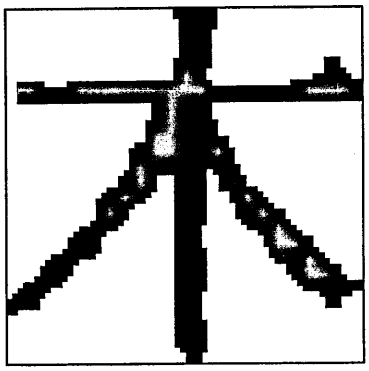
  なお、次元数は方向線素特徴量と同じ 196 次元である。



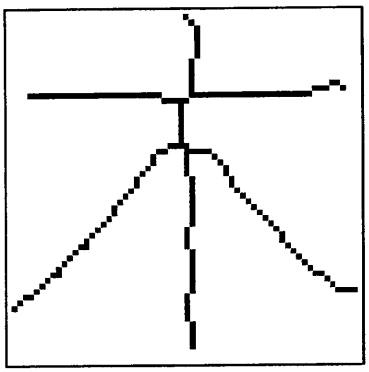
(a) 入力画像



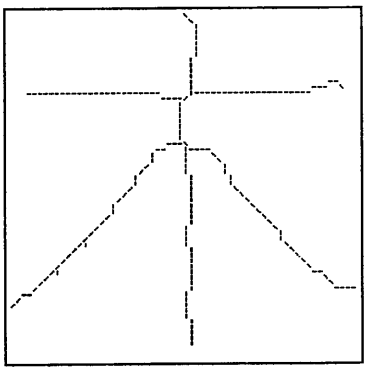
(b) ノイズ除去・スムージング



(c) 正規化

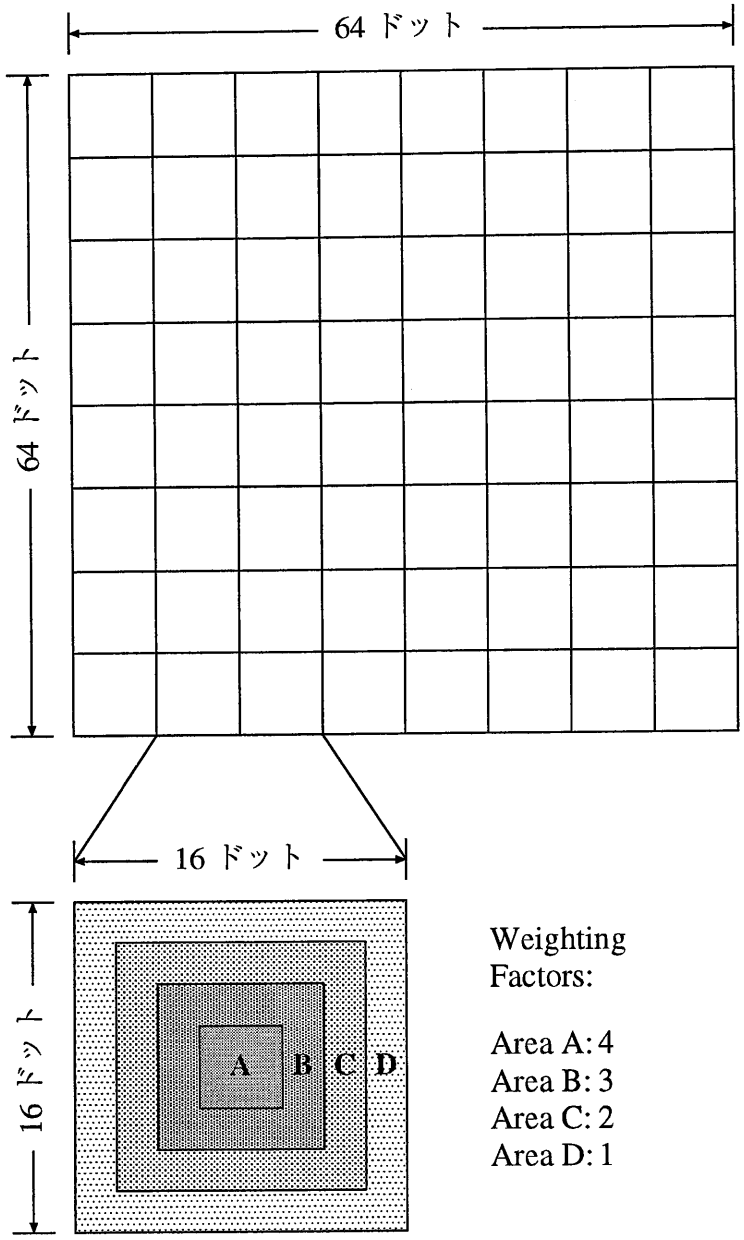


(d) 細線化



(e) 方向線素化

図 2.2: 前処理



方向線素: | ― ＼ ／

方向線素特徴量: $V_m = (V_{m_1}, V_{m_2}, V_{m_3}, V_{m_4})$
where $m = 1 \sim 49$

小領域の数: 49

方向線素特徴量の次元数: 196 (4×49)

図 2.3: 方向線素特徴量

2.4　認識

ここでは最も一般的な全数整合法を説明する。全数整合法は、各文字の辞書ベクトルと特徴抽出で得られた未知入力文字の特徴量で評価値 (距離) を求め、小さいものから順に認識結果とする方法である。アルゴリズムの簡潔性やある程度の高い認識率を得られることから、一般的に用いられている。認識で用いられる辞書・評価値について以下で説明する。

2.4.1　辞書

辞書ベクトルを作成するとき、各字種ごとに、あらかじめ多数の学習パターンを用意しておき、そのパターンから求めた特徴ベクトルの平均を辞書ベクトルとする。シングルテンプレート法、マルチテンプレート法の辞書の構成はそれぞれ次のようになる。

- シングルテンプレート法：

  図 2.4-a に示すように、各字種ごとに一つの辞書を用意する。辞書のサイズは字種数分でよいが、マルチフォントの活字や手書き文字を認識する場合、認識率が低い。

- マルチテンプレート法：

  図 2.4-b に示すように、一つの文字に対して、複数の辞書を用意する。マルチフォントの活字の認識に有効であるが、辞書のカテゴリ数が大きくなる。

2.4.2　評価値

認識を行う際、近さの尺度 (評価値) を定義する必要がある。例えばユークリッド距離 (実際はその 2 乗) を評価値とする場合、字種 kの辞書ベクトルを

$$\boldsymbol{m}^k = (m_1^k, m_2^k, ..., m_N^k) \tag{2.1}$$

とし、入力パターンから求めた特徴ベクトルを

$$\boldsymbol{v} = (v_1, v_2, ..., v_N) \tag{2.2}$$

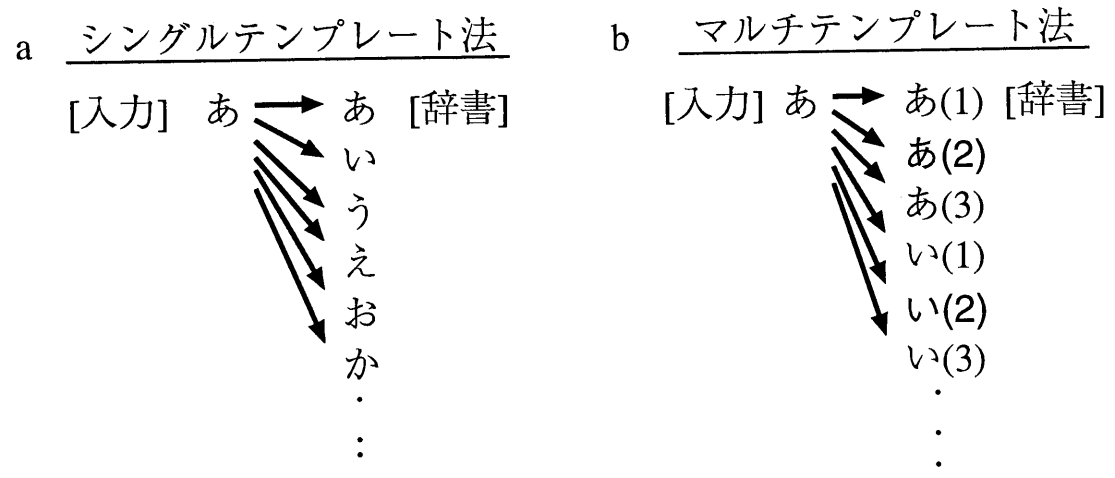


図 2.4: シングルテンプレート法とマルチテンプレート法

とすると、字種 k の評価値 E_k は、

$$E_k = \sum_{i=1}^{N}(v_i - m_i^k)^2 \tag{2.3}$$

となる。E_kの小さいものから候補字種とする (N は特徴ベクトルの次元数で、ここでは $N = 196$)。

2.5　候補出力

評価値計算をした後評価値の小さい順に、第 1 位候補、第 2 位候補、… 、第 n 候補を出力する。

第 3 章

文字特徴量の分布

有効な文字認識手法を検討するために、文字特徴量の空間上での分布等の統計的性質を調べ、文字パターンの分布状況を明らかにする必要がある。そして、誤認識が生じるのがどのような分布のときなのかを把握する。

まず、主成分分析法を用いて異なる字種のサンプルパターン間の分布 (カテゴリ間分布) と同一字種のサンプルパターンの分布 (カテゴリ内分布) を調べた。次に、領域半径の概念を用いて、カテゴリ間の重なりを調査した。調査対象としたのは、数字・アルファベット・平仮名・片仮名計 198 字種である。(人間でも区別できない大文字の「あ」と小文字の「ぁ」などは同一字種とみなした)。具体的な字種は後に表 4.1で示す。これらは特にフォントによる変形が大きく、また文書中での使用頻度の高く、マルチテンプレート化の効果が大きいと期待される字種である。しかも、濁点の有無、濁点と半濁点の違いなどによる類似文字の組合せが多く、誤認識の生じやすい集合でもある。サンプルパターンとして、明朝体・ゴシック体の印刷文字をスキャナで読み取ったものを 1 字種あたり平均 123 個 (最少で 105 個) 用意した。

3.1 主成分分析法を用いた分布形状の調査

カテゴリ間の分布を調べるために主成分分析法を用いた。主成分分析法は多くの変量の値を、できるだけ情報の損失なしに、少数個の主成分で代表させる方法である。多次元空

間上で最も分散の大きい方向を第一主成分、第一主成分と直交する方向のうち最も分散の大きい方向を第二主成分とする。以下、同様に、第三、第四 … 、と定義される。パターンの分散共分散行列を V とするとき、V の固有ベクトルのうち固有値が最大のものが第一主成分となる。次式を満たす λ が固有値、$\boldsymbol{a}$ が固有ベクトルとなる。

$$V\boldsymbol{a} = \lambda\boldsymbol{a}$$

第 n 固有値を λ_n としたとき、第 n 主成分の寄与率は、次のように定義される。

$$\text{第 } n \text{ 主成分の寄与率} = \frac{\lambda_n}{\sum_{i=1}^{N} \lambda_i}$$

ここで、N は次元数であり、方向線素特徴量では $N = 196$ である。第 n 主成分の寄与率は第 n 主成分軸上での分散を表す。

カテゴリ間分布を調べるために、まず調査対象とした各字種の特徴量の重心の集合を主成分分析し、重心ベクトルおよび個々のサンプルパターンをプロットすることで分布形状を確認した。横軸を第一主成分、縦軸を第二主成分としてプロットした結果を図 3.1に、この場合の累積寄与率を図 3.2に示す。さらに、対象字種のうち「あ」「い」「ア」「イ」「0」「A」の 6 字種について、図 3.1と同じ軸上にカテゴリに含まれるすべてのサンプルパターンをプロットした結果を図 3.3 に示す。これらのサンプルパターンのうち 4 個については実際の文字画像も示してある。「い」のサンプルに非常に近い「0」のサンプルがあることがわかるが、実際の画像を見ると、この場合「0」の上と下がかけているため、かなり「い」に近い画像になっていることが分かる。図 3.1、図 3.3で完全に分布状況が把握できるとは言い難いが、カテゴリ内の分布の広がりに対して、カテゴリ間の分布はかなり密であること、第一主成分の固有ベクトルの方向は字種ごとに異なることが分かる。

次に、カテゴリ内の分布がどの程度偏っているかを調べるために、各カテゴリごとに第一主成分の寄与率を求めた。その結果を図 3.4に示す。また、寄与率に関する度数分布を図 3.5に示す。図 3.5より、第一主成分の寄与率は 20 ～ 30%のものが一番多く、高いものは 70%にも達する。196 次元のうち第一主成分だけで 20 ～ 30%の寄与率となることから、かなり偏った分布をしていることが分かる。

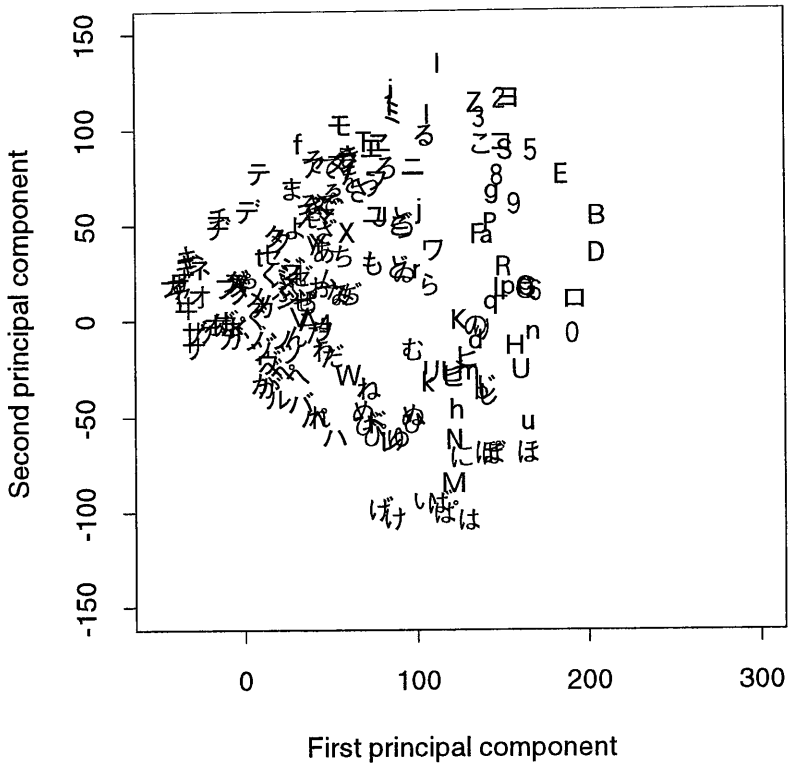


図 3.1: 重心ベクトルの主成分分析の結果 (全字種)

3.2　領域半径の概念を用いた特徴領域の調査

誤認識が生じる場合についてより詳しく分析するためには、各字種の特徴量が分布する特徴領域の相互関係を正しく把握する必要がある。そこで、領域半径 [2] の概念を用いて各字種の特徴領域の重なりの有無を調べた。

ある距離尺度 Eのもとで字種 c の領域半径を

$$r_c = \max_{P \in \mathcal{K}(c)} E(\boldsymbol{m}^c, F(P)) \tag{3.1}$$

と定める (図 3.6(a) 参照)。ここで $\boldsymbol{m}^c$は字種 c の特徴量の重心である。また、$\mathcal{K}(c)$ は字種 c の文字画像の集合であり、Fは文字画像から特徴量への写像を表す。すなわち、$F(P)$ は

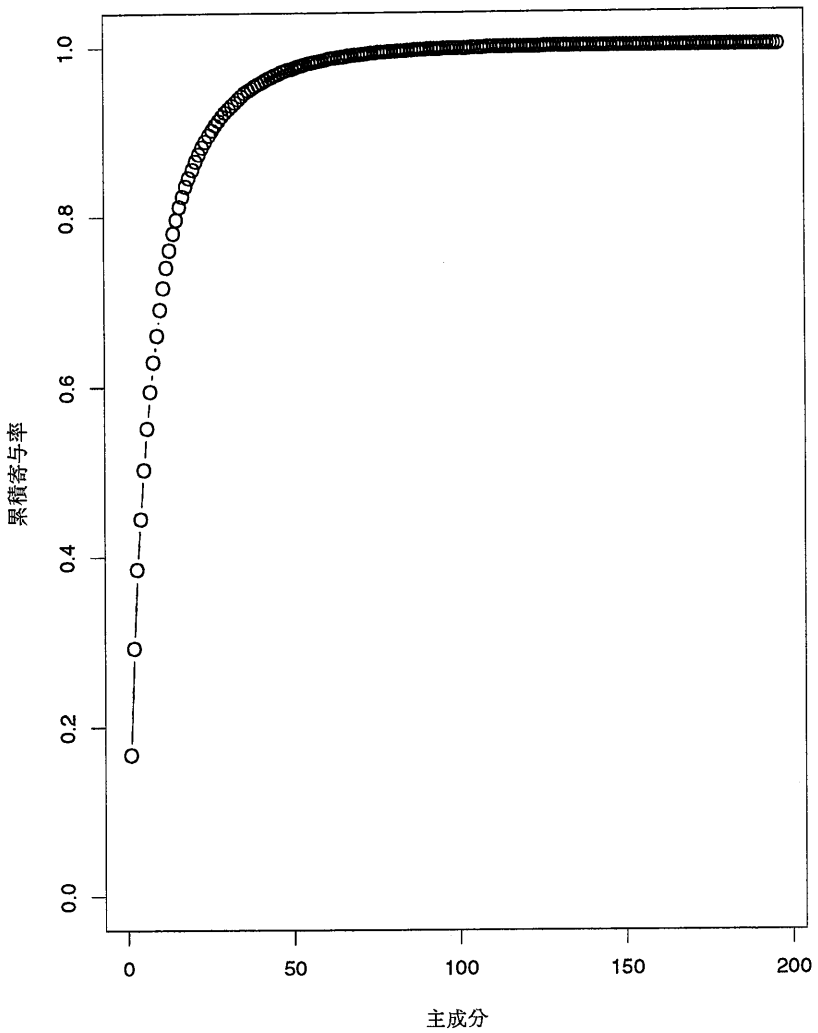


図 3.2: 累積寄与率

字種 c の (実際にはサンプルから求めた) 特徴量を表す。特徴領域を

$$\mathcal{C}(c) = \{\boldsymbol{u}|E(\boldsymbol{m}^c, \boldsymbol{u}) \leq r_c\} \tag{3.2}$$

と定義すれば、次の式 (3.3) の条件が満たされることは字種 c_1 と字種 c_2 の特徴領域が距離尺度 Eのもとで重なりがないことを意味する。

$$\mathcal{C}(c_1) \cap \mathcal{C}(c_2) \neq \phi \tag{3.3}$$

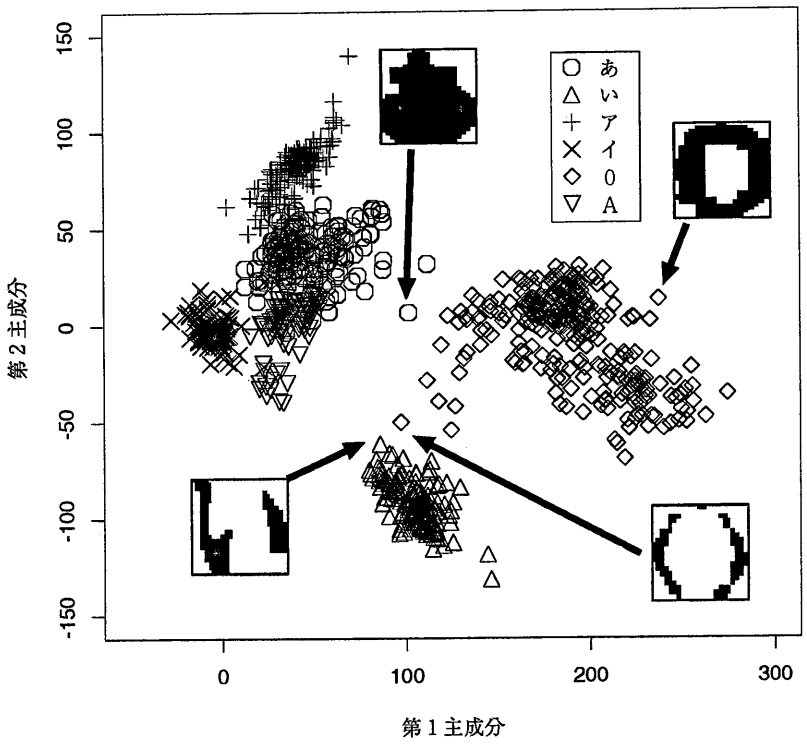


図 3.3: 6 字種のパターンの分布

字種 c_1と字種 c_2の重心間の距離を d とすると、式 (3.3) は $r_{c_1} + r_{c_2} < d$ に等価である。従って、サンプルデータのみを用いたとき、$r_{c_1} + r_{c_2} \geq d$ であれば、字種 c_1と字種 c_2の特徴領域は重なっていると判断される (図 3.6(b) 参照)。

上で述べた考え方をもとに、つぎの 3 種類の距離尺度を用いて重なり具合を調べた。

(a) ユークリッド距離

(b) マハラノビス距離 (分散共分散行列は全字種の平均)

(c) マハラノビス距離 (分散共分散行列は対象とする 2 字種の平均)

全字種対のうち重なっていない対の割合 (分離率) を表 3.1に示す。

ユークリッド距離を距離尺度とした場合、特徴領域は超球となり、マハラノビス距離を距離尺度とした場合は分散の大きい方向に広がった超楕円に近い形となる。表 3.1より、方

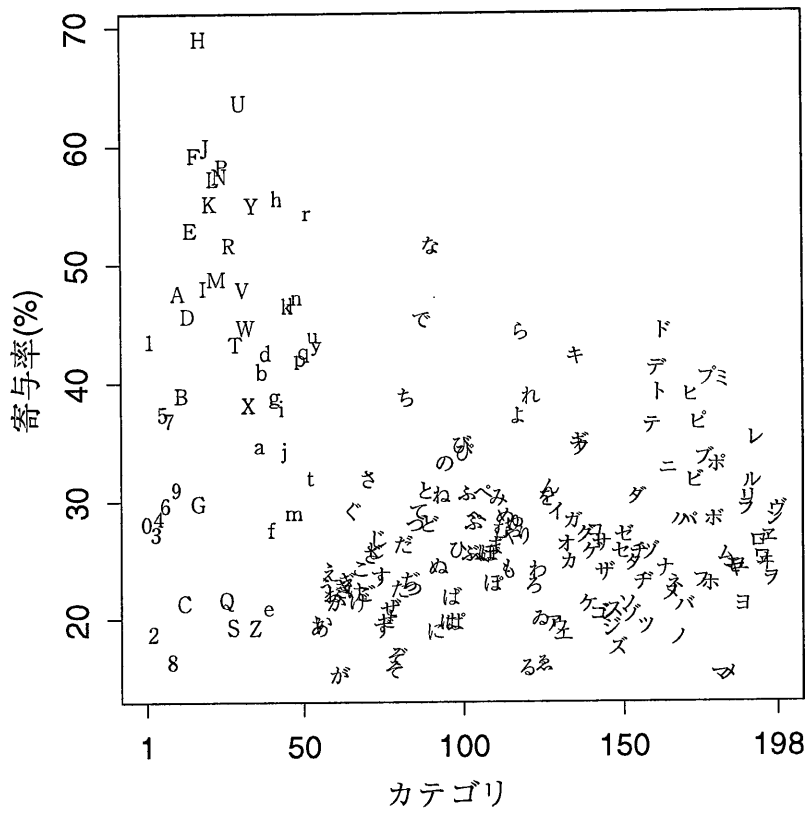


図 3.4: 各字種の第一主成分の寄与率

向線素特徴量を用いた場合の文字特徴量の分布はつぎのようであると考えられる。すなわち、(a) と (c) との比較から、特定の方向の分散のみが大きく、実際に領域が重なっている字種対は少ないことが予想される。また、(b) と (c) との比較から、字種ごとに主成分の方向が大きく異なることが予想される。

3.3　印刷文字と手書き文字の違い

以上の結果はマルチフォント印刷文字を対象としたものであるが、手書き文字を対象とした場合にも同様の結果が得られることを確認している。しかし、マルチフォント印刷文字と手書き文字では分布形状に若干の差がある。例えば、図 3.7と図 3.8は、それぞれ上記調査に用いた印刷文字集合および手書き文字データベース ETL9B の「あ」の特徴量の第

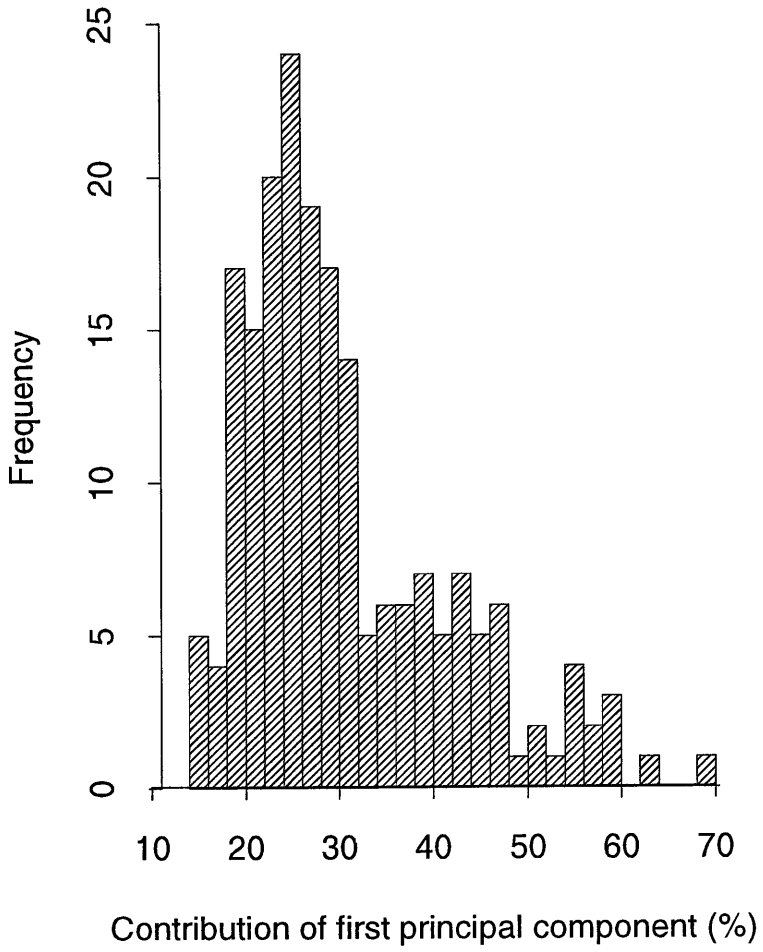


図 3.5: 寄与率に関する度数分布

一主成分軸上での分布を表している。これらを比較して分かるように、マルチフォント印刷文字の方は軸上での分散は小さいが分布が単純な正規分布とは異なりピークが複数存在する。これは、印刷文字にはさまざまなフォントがあり、フォントが同じであれば文字パターンの変動は小さいが、フォントが異なると形状が大きく異なることを意味する。一方手書き文字の方は軸上での分散は大きいが正規分布に近い分布となっている。これは、一

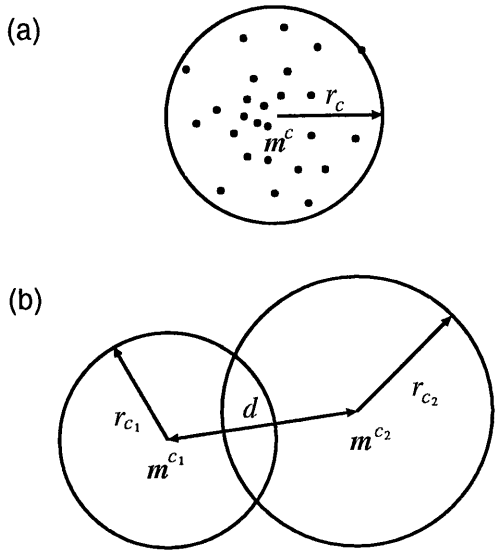


図 3.6: 領域半径と特徴領域

距離尺度	ユークリッド距離	マハラノビス距離	
		全字種の平均	2 字種の平均
重なっている	19474	19503	284
重なっていない	29	0	19219
重なっていない率	0.15%	0%	98.54%

カテゴリ対の総数：19503

表 3.1: 重なり具合の調査結果

つ一つの文字パターンの変動は大きいがその分布はある平均パターンのまわりに均等に分布していることを意味する。

3.4　調査結果についての考察

文字特徴量の分布状況を調べる手法を述べ、数字・アルファベット・平仮名・片仮名を用いて分布状況を調査した。これらの結果から、方向線素特徴量を用いた場合の文字特徴量の分布は以下のようになっていると考えられる。すなわち、主成分分析を用いた調査から、同一字種の特徴量の分布の広がりに対して異なる字種の特徴量間の分布はかなり密である

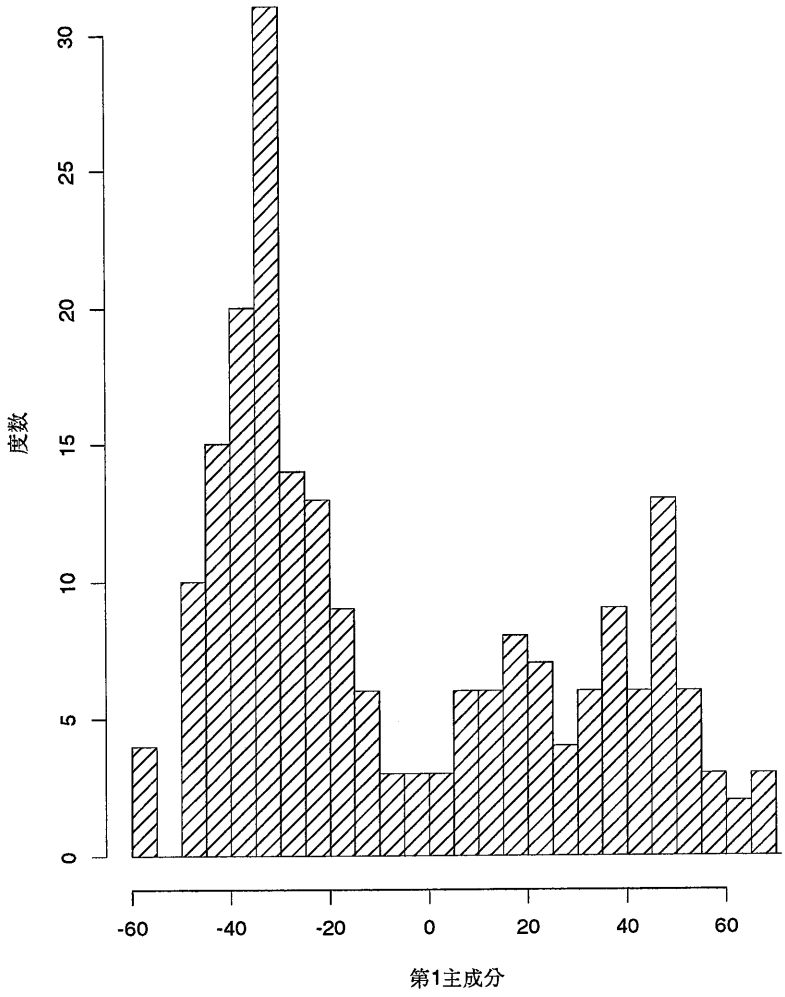


図 3.7: 文字「あ」の特徴量の第 1 主成分の分布 (印刷文字)

こと、字種内の分布はかなり偏っていることが分かった。領域半径の概念を用いた重なり具合の調査から、字種内の分布は特定の方向の分散のみが大きく、実際に領域が重なっている字種対は少ないこと、字種ごとに主成分の方向が大きく異なることが分かった。

さらに、マルチフォント印刷文字と手書き文字の分布形状の違いについての調査から、マルチフォント印刷文字認識には評価値としてユークリッド距離を用いて複数のテンプレートを用意するマルチテンプレート法、手書き文字認識にはパターンの分布形状をよく表す

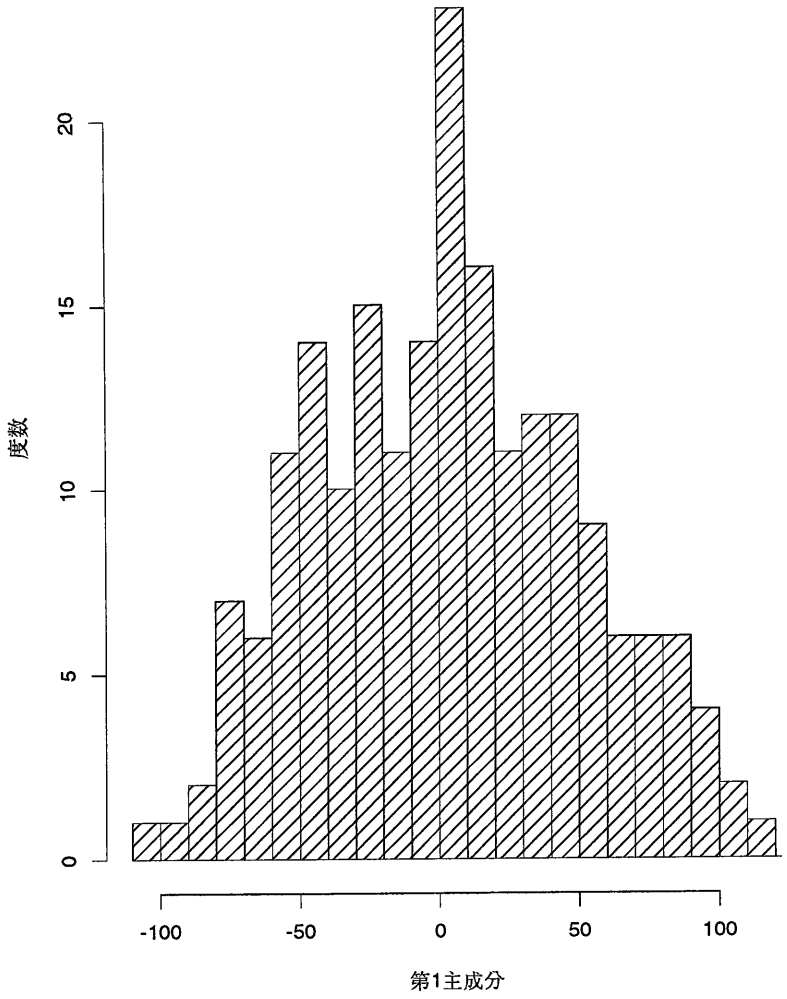


図 3.8: 文字「あ」の特徴量の第 1 主成分の分布 (手書き文字)

ような評価値を用いた識別が有効であると考えられる。以下の章ではそれぞれについて有効な認識手法を提案する。

第 4 章

カテゴリ間の分布を考慮した文字認識用マルチテンプレート辞書の構成法

— 字種適応クラスタリング法 —

4.1 はじめに

文字認識において辞書のマルチテンプレート化は認識の高精度化の一つの手段であるが、辞書が大きくなり記憶容量・計算時間が増加する。また、同一字種のテンプレートを複数持つことは異なる字種のテンプレート間の距離が近くなることを意味し、逆に誤認識となる字種も現れる等の問題がある。これらの問題を解決するための、マルチフォント印刷文字認識のための文字特徴量の分布形状を考慮した辞書作成法について検討する。具体的には、前章の考察結果をもとに、同一字種のサンプルパターンの分布 (カテゴリ内分布) と異なる字種のサンプルパターン間の分布 (カテゴリ間分布) を考慮することで誤認識が生じ得る状況を推定し、誤認識の可能性のある字種のテンプレート数を増やす操作を繰り返すことで字種ごとにテンプレート数の異なる辞書を人が介在せずに作成する手法 (字種適応クラスタリング法) を提案する。

文字認識に用いる距離尺度としては、ユークリッド距離、マハラノビス距離などさまざまなものが提案されている。これらのうち、マハラノビス距離は多次元正規分布の確率密

度関数の式から導かれ、文字特徴量が正規分布をしている場合にはその分布状況をよく表す距離尺度である。しかし計算コストが大きく、サンプル数が少ないと分散共分散行列が求まらずマハラノビス距離そのものが定義できない等の問題がある。ユークリッド距離の場合はマハラノビスと比べて計算量が少なく、サンプルが少ないことによる悪影響が比較的少ない。したがって、本章では距離尺度としてユークリッド距離を用い、認識を高速・高精度に行うマルチテンプレート辞書の構成法について検討する。

字種適応クラスタリング法は、まず字種ごとのサンプルパターンの集合をそれぞれ一つのクラスタとみなし、誤認識の可能性のあるクラスタを含むカテゴリを分割してクラスタ数を増やしていく手法である。まずカテゴリを分割する条件について述べ、アルゴリズムを説明する。

4.2　カテゴリを分割する条件

ユークリッド距離を距離尺度とした場合に誤認識が生じる典型的な例として、図 4.1(a) のように特徴量が分布している場合が考えられる。図の 2 個の楕円 j,kはそれぞれ、文字 c_1と文字 c_2 の一つのクラスタであり、$\boldsymbol{m}^j$、$\boldsymbol{m}^k$はそれぞれの重心である。文字 c_2のサンプルパターンの 1 つである点$\boldsymbol{x}$に着目する。図からわかるように、点$\boldsymbol{x}$から文字 c_1の重心との距離 d_jは文字 c_2の重心までの距離 d_kより短いので、ユークリッド距離を評価値として認識を行うと、点$\boldsymbol{x}$は文字 c_1に誤認識される。このような誤認識を避けるためのカテゴリを分割する条件について考える。

クラスタ kの重心を$\boldsymbol{m}^k$、第 i 主成分の固有値をλ_i^k 固有ベクトル (単位ベクトル) を $\boldsymbol{a}_i^k$とする。固有値は主成分軸上での分散を表すから、適当な定数αを用いて、次の 2 点でこの特徴領域の第 i 主成分軸の両端を表せる。

$$\boldsymbol{m}^k + \alpha\sqrt{\lambda_i^k}\boldsymbol{a}_i^k \tag{4.1}$$

$$\boldsymbol{m}^k - \alpha\sqrt{\lambda_i^k}\boldsymbol{a}_i^k \tag{4.2}$$

ここでは$\alpha = 3.3$ とした。これは、軸上の分布が正規分布である場合に、クラスタに属するパターンの 99.9%が含まれる値である。

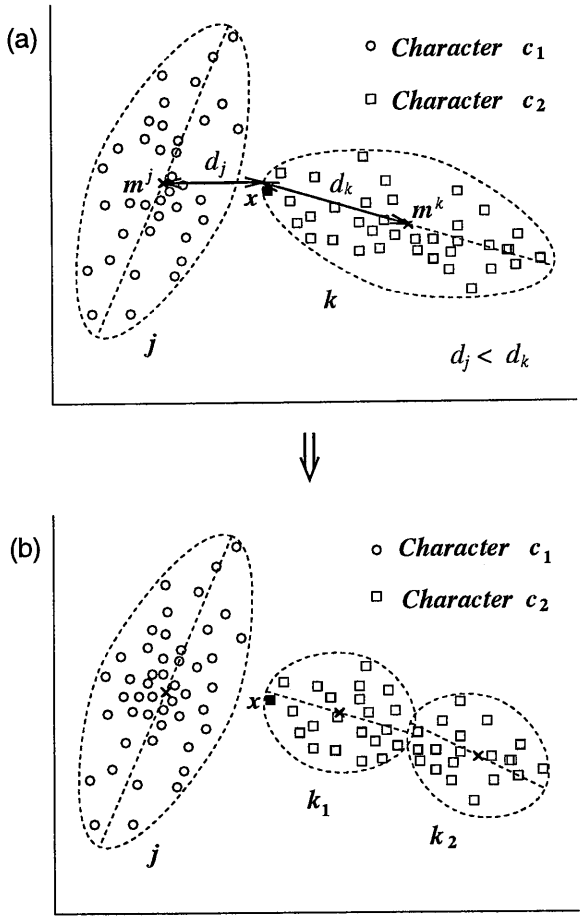


図 4.1: 分割する条件

第 L 主成分軸まで考慮することにすれば、ある $i(1 \leq i \leq L)$ に対して次の式 (4.3)、式 (4.4) が成り立つとき、クラスタ kの第 i 主成分軸を、「誤認識の可能性のある軸」と定義し、クラスタ kを含むカテゴリを分割する必要があるとみなす。

$$\min_{j \in \bar{J}(c)} E(\boldsymbol{m}^k \pm \alpha\sqrt{\lambda_i^k}\boldsymbol{a}_i^k, \boldsymbol{m}^j) \ < \ \alpha\sqrt{\lambda_i^k} \tag{4.3}$$

$$N(c) < M \tag{4.4}$$

但し、c はクラスタ kの字種であり、$\bar{J}(c)$ は c 以外のすべての字種のクラスタの集合で、$N(c)$ は字種 c を構成するクラスタの数である。Mは定数であり、一つのカテゴリが必要以上に細かく分割されるのを防ぐために設定した。図 4.1を用いて具体的に説明する。図

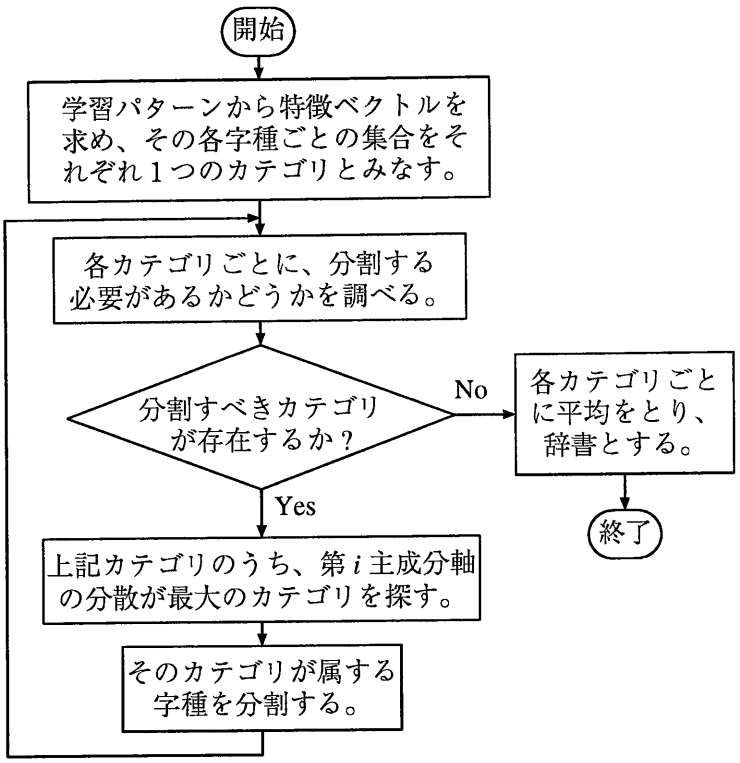


図 4.2: 字種適応クラスタリング法のアルゴリズム

(a) で、

$$d_j = E(\boldsymbol{m}^k + \alpha\sqrt{\lambda_1^k}\boldsymbol{a}_1^k, \boldsymbol{m}^j) \tag{4.5}$$

$$d_k = \alpha\sqrt{\lambda_1^k} \tag{4.6}$$

である。前で述べたように、(a) では点$\boldsymbol{x}$は文字 c_1に誤認識されるが、この場合、$d_j < d_k$ であり、式 (4.3) が満たされるから、クラスタ kは分割される。図 (b) のように kが k_1と k_2 に分割されると、同じ点$\boldsymbol{x}$からクラスタ k_1の重心までの距離の方が短くなり、正しく文字 c_2に認識されるようになると考えられる。

4.3　字種適応クラスタリング法のアルゴリズム

字種適応クラスタリング法のアルゴリズムを図 4.2に示す。

まず、各字種のサンプルパターンから特徴量 (ベクトル) を求め、各字種ごとの特徴量の集合をそれぞれ一つのクラスタとみなす。各クラスタごとに、式 (4.3)、(4.4) を満たす軸が存在するかどうかを調べる。そして、式 (4.3) を満たす (k,i) のうち、λ_i^kの値が最も大きいもの[1]を探し、そのクラスタを含む字種のクラスタ数を 1 増加する[2]。この操作を式 (4.3)、(4.4) を満たすクラスタがなくなるまで繰り返す。そして、最終的に得られた各クラスタごとに重心を求め、辞書を構成する。分割手法としては K-means 法を用いた。kが字種 c のクラスタであるとき、再分割のための初期クラスタ中心は、$\boldsymbol{m}^k \pm \varepsilon \boldsymbol{a}_i^k$と、$k$以外の c のクラスタの重心 (存在すれば) である (εは微小長さ)。

具体例を図 4.3 に示す。図 4.3の (a) のクラスタ k_1が第一主成分を考慮したときに分割すると判断された場合には、初期クラスタ中心をクラスタ k_1の重心から k_1の第一主成分の方向に$\pm\varepsilon$だけ移動した 2 点とし、すべての特徴ベクトルを、K-means 法を用いてクラスタリングする。その結果は (b) のようになる。以下、(b) のクラスタ k_2の第一主成分、(c) のクラスタ k_2の第二主成分が上記判断法により選択されたとすると、最終的に (d) のようなクラスタ構成となる。

4.4　認識実験

4.4.1　予備認識実験

字種適応クラスタリング法 (以下、本手法とよぶ) では、パラメータとして主成分軸の数 L と 1 カテゴリあたりの最大クラスタ数 M が必要となる。適当なパラメータを設定するために、**第 3章**で調査対象とした 198 字種を用いた予備認識実験を行った。調査対象としたデータをそのまま学習データとし、各クラスタの第一主成分のみを考慮したもの ($L=1$ の場合) と第三主成分まで考慮したもの ($L=3$ の場合) の 2 種類について、式 (4.4) の M を 1 から 8 まで変化させて認識用辞書を作成した ($M=1$ の場合はシングルテンプレート

[1]分割するクラスタを選択する基準としては、クラスタに含まれるサンプル数が最も多いもの、クラスタの第一主成分の固有値が最も大きいもの等についても検討したが、実験により本手法が最も良い認識率が得られることが分かった。

[2]対象クラスタを 2 分割するのではなく、カテゴリ全体を再分割する。前者も検討したが、後者の方が良い認識率が得られることが分かった。

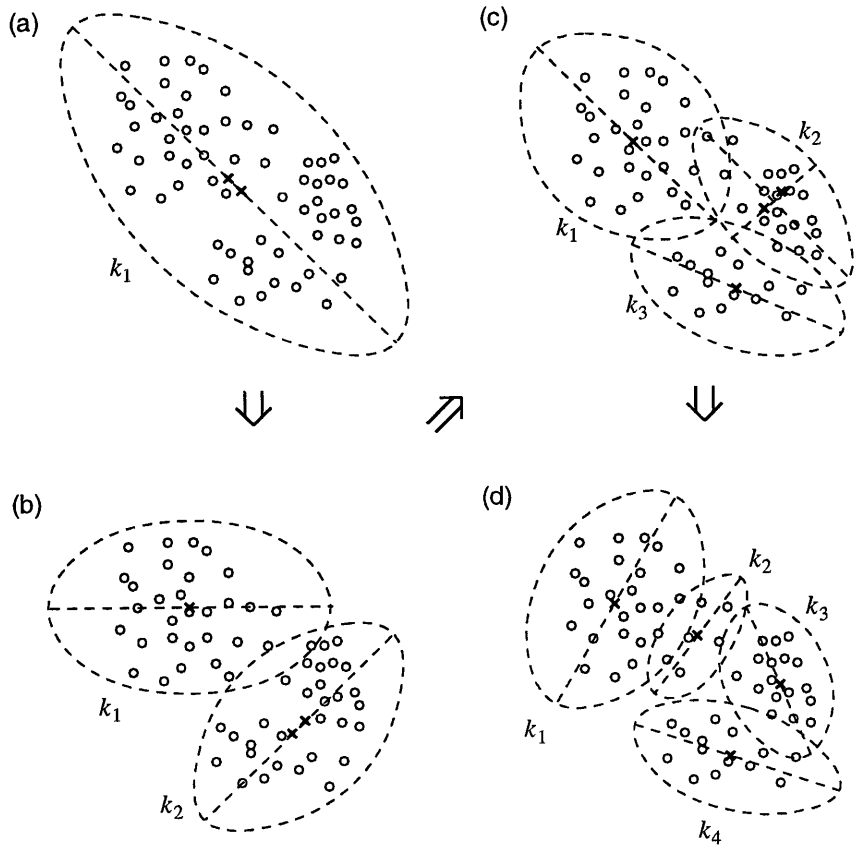


図 4.3: 初期クラスタ中心の取り方

の辞書となる)。そして、学習データとは別のフォントの明朝体とゴシック体を各 2 セットずつ評価用のデータとした。評価値としてはユークリッド距離を用いた。

さらに、本手法では、クラスタの端点を取るとき、サンプルの軸上の分布を正規分布と仮定し、クラスタに属するパターンの 99.9%が含まれるように、$\alpha = 3.3$ とした。図 4.4 は「あ」のすべてのサンプルについて主成分分析し、第一主成分、第二主成分と第三主成分の各主成分軸上にプロットした様子を示した。図の中心付近の点は重心であり、周辺の 4 点はそれぞれの主成分に対応する端点を表す。図からわかるように、特徴領域は標準偏差の 3.3 ($\alpha = 3.3$) 倍とした場合には、若干はみだすところもあるが、サンプルパターンの分布領域をよくカバーしていると言える。

また、カテゴリ間の分布を考慮することの効果を確認するため、本手法で用いている K-平均法を階層的に適用する手法である LBG 法 [14] を用い、字種ごとにテンプレート数固

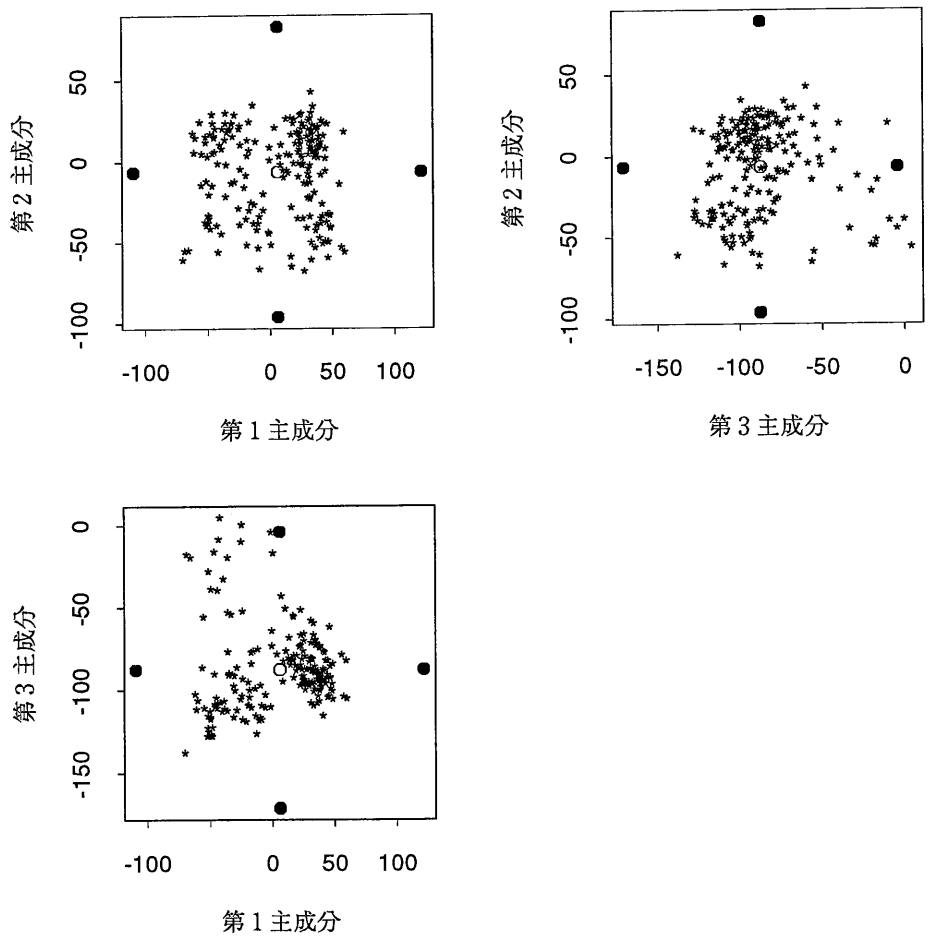


図 4.4: サンプルパターンと重心・端点 (文字「あ」)

定で作成した辞書を用いた認識実験 (以下、従来法とよぶ) も行った。1 字種あたりのテンプレート数は 1、2、4、8 とした。

1 字種あたりの平均テンプレート数に対する認識率 (1 位認識率) を図 4.5に示す。図 4.5 で、本手法の各点は、平均テンプレート数が小さい順に式 (4.4) の M の値が 1 から 8 の場合に対応する。図 4.5より、テンプレート数が同程度の場合、本手法で作成した辞書を用いて認識した方が $L = 1$、$L = 3$ の場合ともに従来法よりも良い結果が得られることが分かる。また、従来法ではテンプレート数が 4 の場合に認識率が 96.5%でピークとなっている。本手法で $L = 3$ の場合は $M = 3$ で認識率がほぼ頭打ちになっているが、$M = 3$ で従来法を超える認識率 (96.6%) が得られており、しかもこの場合の 1 字種あたりの平均テン

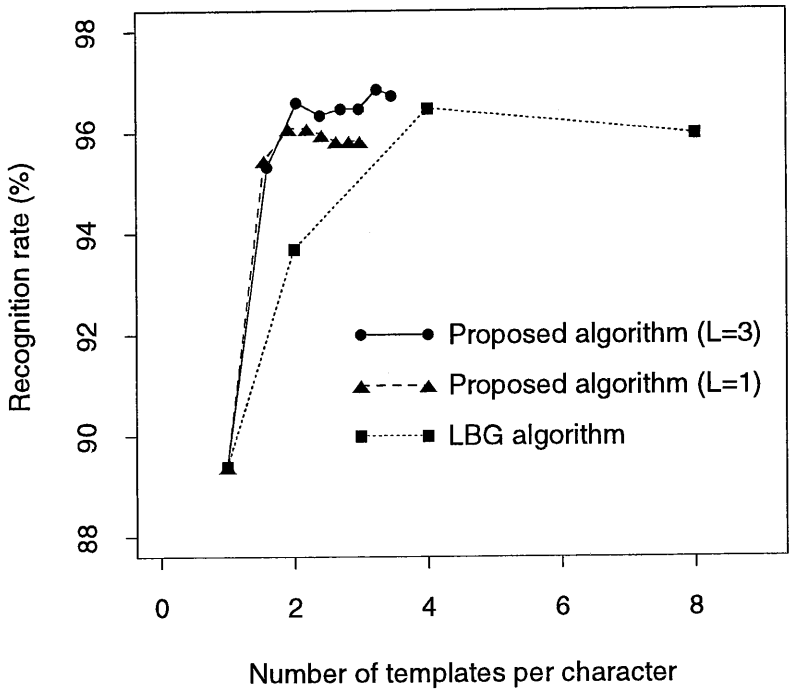


図 4.5: 予備実験結果

プレート数は 2.0 で、従来法の 2 分の 1 である。サンプルとして用意した明朝体とゴシック体の 2 種類のフォントに対し、$M \geq 3$ の場合に $M = 2$ の場合よりも良い認識率が得られているのは、同じ明朝体、あるいはゴシック体でも形が大きく異なり分割すべきクラスタが存在するからである。同じ明朝体でもフォントの形状の異なる字種の例として、「で」の 2 種類のサンプル (大きさの正規化をほどこしたもの) を図 4.6に示す。これらは $M = 3$ で異なるクラスタに属したものであるが、濁点の位置が大きく異なり、特に正規化した場合にその形状が大きく違っていることが分かる。

本手法による辞書の構成例として、$M = 3$ の場合の字種ごとのテンプレート数を表 4.1 に示す。また、第三主成分まで考慮した場合、式 (4.3) を満たすクラスタと主成分軸の組 (k, i) のうち固有値λ_i^kの値が最大となるクラスタ kの第 i 主成分軸を選択して初期クラスタ中心を設定するが、分割の過程で各主成分軸が選択された回数を表 4.2に示す。また、評価

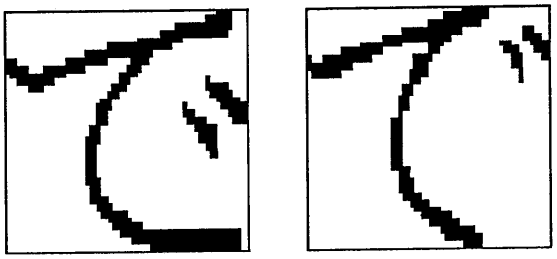


図 4.6: 「で」のサンプルの例

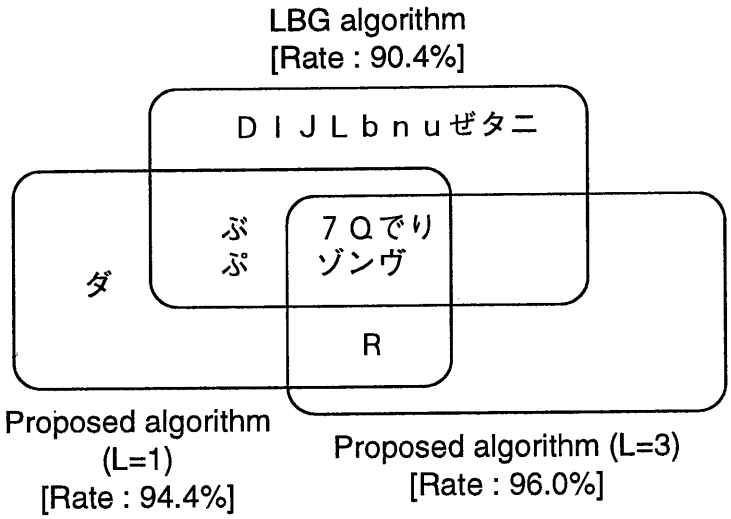


図 4.7: 誤認識字種の例

用の 4 セットのデータのうち、シングルテンプレートの辞書を用いて認識したときに最も認識率が低かった 1 セット (認識率 81.3%) について、上記の各手法で誤認識となった字種を図 4.7に示す。但し、総テンプレート数が同程度の辞書で比較するために、従来法ではテンプレート数が 1 字種あたり 2 個のもの、本手法では $M = 3$ のもの (平均テンプレート数は $L = 1$ のとき 1.92、$L = 3$ のとき 2.05) を用いた。それぞれの辞書での認識率も [ ] 内に示してある。

表 4.2より、選択された主成分は第一主成分がほとんどで、第二主成分は第一主成分の 10 分の 1 程度しか選択されず、第三主成分はほとんど選択されないことが分かる。従って、本手法では第一主成分が非常に重要な役割を果たす。但し、第一主成分のみでは不十分である。例えば、図 4.7から分かるように、「ぶ」と「ぷ」の 2 字種は第一主成分のみを考慮

表 4.1: 字種ごとのテンプレート数

字種	テンプレート数	
	$L=1$	$L=3$
2 3 4 5 8 G S W Z あいえおくぐけこごせぜそぞたつづにのひまみむめやゆよるれゐゑをんアイウケゲコゴサザジスズゼナヌネノハバパフマミムメモヤヨラルレヲ	1	1
げ	2	
うざちね	1	2
0 6 7 A C M R T V Y d e j m y すずだてなぬはへほもらろエオカキシセヒユヰヴ	2	
きぎさぢふぶぷわ	1	3
X しじダヱ	2	
1 9 B D E F H I J K L N P Q U a b f g h i k l n p q r t u かがでとどばぱびぴべぺぼぽりガギクグソゾタチヂツヅテデトドニビピブプホボポリロワン	3	

表 4.2: 各主成分軸が選択された回数

第一主成分	第二主成分	第三主成分
185	21	1

した手法では誤認識であるが、第三主成分まで考慮した場合は正しく認識されている。表 4.1より、これらの字種は第一主成分のみ考慮した場合はテンプレート数が 1 であるが、第三主成分まで考慮した場合はテンプレート数が 3 となって正しく認識されるようになった、つまり、分割する必要があった字種であると思われる。「ぶ」と「ぷ」の第一主成分軸のなす角をθとし、これら 2 字種の固有ベクトルを用いて$\cos\theta$の値を求めると 0.94 となり、ほぼ平行であることが分かった。このような場合、第一主成分のみを考慮すると、重心どう

しが近くても式 (4.3) の条件が満たされず、分割する必要がないと判断され、テンプレート数は 1 となる。この場合は第二主成分を考慮することで分割する必要があると判断され、テンプレート数は 3 となって、正しく認識されるようになった。この例のような場合があるため、第二主成分以降も考慮する必要がある。

また、$L = 4$ とした実験でも第四主成分が選択されることはなかったので、第四主成分以降はほとんど無視できると思われる。

4.4.2　認識実験

予備実験の結果を踏まえ、JIS 第一水準の漢字 2965 字種を含め、計 3163 字種に対する本手法の有効性を確認する。ここでは、明朝体 12 種類、ゴシック体 12 種類の計 24 種類のフォントを印刷したものをスキャナで読み取り、認識実験用データとして用意した。これらは異なるプリンタから出力したか、同一プリンタの場合はフォントが異なるため、すべてが別のフォントであると言える。明朝体とゴシック体 1 種類ずつを 1 つのセットとし、データを 12 個のセットに分ける。ローテーション法を用い、12 セットのうち 1 セットを除いた 11 セット (22 種類) のフォントを用いて認識用辞書を作成し、除いた 1 セット (2 種類) を評価用データとして認識する。この操作を各セットについて行い、辞書のテンプレート数、認識率について平均を求めた。分割するかどうかの判断基準、分割手法に関しては予備実験と同じものを用いた。また、予備実験における考察から本手法においては $L = 3$、$M = 3$ とし、従来法として LBG 法における 1 字種あたりのテンプレート数が 4 の場合、およびシングルテンプレートの辞書を用いた場合と比較した。

実験の結果を表 4.3に示す。本手法および従来法で辞書を作成した場合の辞書の 1 字種あたりの平均テンプレート数、およびその辞書を用いて認識した場合の平均認識率を示してある。シングルテンプレートの辞書による認識結果も合わせて示してある。従来法、本手法ともに認識率は 98.46% とシングルテンプレート法と比較して高い値となっており、辞書のマルチテンプレート化の効果が確認された。また、1 字種あたりの平均テンプレート数を比較すると、従来法が 4 であるのに対して本手法は 1.68 であり、40% 程度に減少している。すなわち、カテゴリ間分布を考慮して必要なもののみテンプレートを複数化することで、認識率を下げずに認識用辞書の総テンプレート数を抑えることができた。認識時

表 4.3: 実験結果

	シングル	マルチテンプレート	
	テンプレート	従来法	本手法
テンプレート数	1	4	1.68
認識率	97.61%	98.46%	98.46%

の計算時間もほぼ辞書のテンプレート数に比例して減少するため、本手法により計算時間・記憶容量ともに減らすことが可能となる。JIS 第一水準の漢字を含めた全字種を対象とした場合にも、本手法の有効性が確認された。

4.5　まとめ

本章では、文字認識においてマルチテンプレートの辞書を用いる際の従来の問題点を指摘し、これを解決するための方策について検討した。まず、前章で調査した空間上での特徴量の分布状況から推定される誤認識が生じる典型的な状況について述べた。それをもとに、誤認識の可能性のある字種のテンプレート数を増やす操作を繰り返すことで、総テンプレート数を抑えたマルチテンプレートの辞書を構成する手法として、字種適応クラスタリング法を提案した。さらに、本手法を用いて認識用辞書を作成して認識実験を行い、本手法で用いている K-means 法を階層的に適用する手法である LBG 法と比較して、認識率を下げずに辞書のテンプレート数を大幅に減少させることができることを確認し、カテゴリ間の分布を考慮することの効果を示した。

文字認識に関する過去の研究では、各字種の適当なテンプレート数の判断法についての検討はほとんどなされていなかった。また、異なる字種の特徴領域間の分布について考慮したアルゴリズムはほとんどなかった。本研究は統計的性質を考慮し、字種ごとのテンプレート数の新しい判断法を提案するものである。

但し、本手法で分割すべきクラスタが選択できるのは異なる字種の特徴領域に重なりがない場合である。重なりがある場合の対処法を考案して本手法に組み込むことは今後の課題である。また、本手法は認識対象や特徴量を制限するものではないので、方向線素特徴量以外の特徴量での有効性を確認することも今後の課題である。

第 5 章

特徴領域の推定による手書き文字の高精度認識法

5.1 はじめに

本章では、手書き文字認識における候補を高精度に選出する手法を提案する。提案する認識アルゴリズムにおいて、まず、学習サンプルにより各字種の文字パターンの特徴量が分布する領域として、一次元特徴領域と多次元特徴領域の二種類の特徴領域を統計的に定める。そして、これらの二種類の特徴領域を同時に参照することで、未知パターンが属している可能性のある特徴領域を推定し、候補選出および得られた候補の信頼性の検証を行う。分布形状を表すパラメータを正確に推定することができれば、多次元特徴領域は文字認識に非常に有効である。しかし、一般に学習サンプルから推定したパラメータには誤差が含まれ、特に学習サンプルが少ない場合においてその誤差がかなり大きく、悪影響がある。一方、一次元特徴領域は認識能力はそれほど高くないが、一次元空間上でのサンプルパターンの分布形状を正しく表すことができる。このような特徴を持つ二種類の特徴領域を同時に参照することで、それぞれの利点を生かし、信頼度の高い手書き文字認識の候補を選出することが可能になる。

多次元特徴領域を定義するために領域半径の概念を用いる。そのため、各字種の文字パターンの特徴量が分布する領域 (特徴領域) を正確に表すことが要求される。そのための距

離尺度としてマハラノビス距離と確率的に同等な評価関数 (距離尺度) となる簡素化マハラノビス距離 (Simplified Mahalanobis Distance; SMD) を提案する。そしてこの新しい距離尺度の振舞いをシミュレーション実験により明らかにする。さらに、提案するアルゴリズムの有効性を手書き文字データベース ETL9B を用いた実験により確認する。

なお、特徴量としては孫らの提案した改良型方向線素特徴量 [12] を用いた。

5.2　領域半径

文字認識の候補の信頼度を測る有効な手段の一つとして、領域半径 [2] の概念がある。**3.2** で述べたように、領域半径の定義式は次のようになる。

$$r_i = \max_{P \in K(i)} E(\boldsymbol{\mu}_i, F(P)) \tag{5.1}$$

未知入力パターンから得られたベクトルと標準パターンとの評価値が r_i と比べて小さいほど字種 i である可能性が高い。ここでは、この概念を用いた認識結果の正しさの検証を取り入れる。

5.3　候補選出のアルゴリズム

3.3 のマルチフォント印刷文字と手書き文字の分布形状の違いについての調査結果から、マルチフォント印刷文字と手書き文字はそれぞれ異なる特徴を持つ。マルチフォント印刷文字の場合は複数のクラスタを組み合わせたような複雑な形状であるのに対し、手書き文字の場合は分布形状は単純な正規分布に近いが分散が大きいという特徴があった。つまり、活字と比べて筆記者の癖による文字パターンの変形が多様で、複雑である。これより、手書き文字認識にはクラスタリングによるテンプレート複数化よりもパターンの分布形状をよく表す評価値を用いる方法が有効であると考えられる。また、従来のいかなる場合にも一個の候補を出力する認識手法より、複数の候補を容認し、リジェクト機能も取り入れて、統計的に見て正しいという保証がある候補を出力することが重要である。後処理とうまく組み合わせることを図る手書き文字認識の特徴を重視し、柔軟性のある認識システムを構築することが可能になる。したがって、本研究では、領域半径の概念および簡素化マハラ

ノビス距離を用い、手書き文字を高精度に認識するアルゴリズムを提案する。本アルゴリズムの特徴は、候補選出の際に、多次元と 1 次元の 2 つの特徴領域を定義し、判断の基準として用いることである。多次元の特徴領域は前述のように学習サンプルが少ない場合において、推定したパラメータには大きな誤差が含まれ、悪影響を与える。一方、1 次元の特徴領域の場合そのような誤差は少ないが、他の次元の情報を無視するために識別能力が低い。したがって、提案する手法は、両者を組み合わせることで互いの欠点を補い、高精度な候補選出を可能にするものである。

未知入力パターンの特徴ベクトル $\boldsymbol{x}$ が与えられた場合、まず特徴ベクトル$\boldsymbol{x}$から各字種 i までの評価値 (距離) $E(\boldsymbol{x}, \boldsymbol{\mu})$ を計算する。そして、暫定的な候補文字集合を

$$\mathrm{Cand}(\boldsymbol{x}) = \left\{ i \left| \frac{E(\boldsymbol{x}, \boldsymbol{\mu}^i)}{r_i} \leq \alpha \right. \right\}, \tag{5.2}$$

と定義する。α は定数である。

$\mathrm{Cand}(\boldsymbol{x})$ は 2 個以上の要素を含む場合が多いので、さらに詳細分類が必要である。$\boldsymbol{x}$ が文字 i のベクトルかどうかを調べるために、文字 i の 1 次元の特徴領域を定義し、これを用いる。$\boldsymbol{v}$ を 文字 i のサンプル集合 $\boldsymbol{S}(i)$ に属している一つのサンプルとする。また、$\boldsymbol{\mu}_i$ と $\boldsymbol{\mu}_j$ はそれぞれ文字 i と文字 j の重心を表す。$\boldsymbol{v} - \boldsymbol{\mu}_i$ の $\boldsymbol{\mu}_i$ と$\boldsymbol{\mu}_j$ を結ぶ直線への射影を $z_{ij}(\boldsymbol{v})$ とすると、

$$z_{ij}(\boldsymbol{v}) = \frac{(\boldsymbol{v} - \boldsymbol{\mu}^i) \cdot (\boldsymbol{\mu}^j - \boldsymbol{\mu}^i)}{\|\boldsymbol{\mu}^j - \boldsymbol{\mu}^i\|}. \tag{5.3}$$

となる。

ここで、$z_{ij}(\boldsymbol{v})$ の標準偏差を s_{ij} とする。$|z_{ij}(\boldsymbol{x})|/s_{ij} \leq \theta$ が成り立つとき、未知入力ベクトル $\boldsymbol{x}$ は文字 i である可能性があると判断できる。すなわち、この $\theta \cdot s_{ij}$ は一次元空間上の領域半径とみなすことができる。θ は定数である。

提案するアルゴリズムを 図 5.1 に示す。まず、各字種 i のサンプルパターンの集合 $\boldsymbol{S}(i)$ が与えられる。そして字種 i の平均ベクトル $\boldsymbol{\mu}^i$ 、領域半径 r_i および字種対 (i, j) の s_{ij} を計算する。さらに、評価関数 E を計算するための各パラメータも求める。これらの値で辞書を構成する。

任意の一つの未知入力パターンのベクトル $\boldsymbol{x}$ に対して、特徴ベクトル $\boldsymbol{x}$ から字種 i の距離を計算し、式 (5.2) により一回目の候補選出を行う。以下の処理は得られた候補の数により異なり、次のようになる。

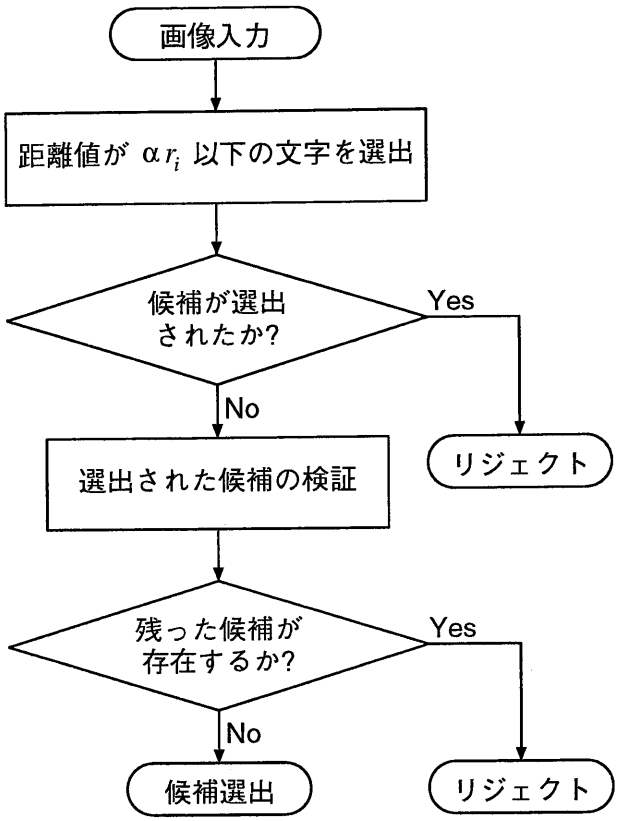


図 5.1: 認識アルゴリズム.

- $|\mathrm{Cand}(\boldsymbol{x})| = 0$ の場合、つまり式 (5.2) を満たす候補が存在しない場合は、認識不可能であるとしてリジェクトする。

- $|\mathrm{Cand}(\boldsymbol{x})| \geq 2$ の場合、$\mathrm{Cand}(\boldsymbol{x})$ の部分集合 $\mathrm{Cand}'(\boldsymbol{x})$ を以下のように定義する。

$$\mathrm{Cand}'(\boldsymbol{x}) = \left\{ i \left| \max_{j \in \mathrm{Cand}(x)-\{i\}} \frac{|z_{ij}(\boldsymbol{x})|}{s_{ij}} \leq \theta \right. \right\}. \tag{5.4}$$

  $|\mathrm{Cand}'(\boldsymbol{x})| > 0$ ならば $\mathrm{Cand}'(\boldsymbol{x})$ を最終候補の集合として出力する。$|\mathrm{Cand}'(\boldsymbol{x})| = 0$ ならば、適当な候補がないとしてリジェクトする。

- $|\mathrm{Cand}(\boldsymbol{x})| = 1$ の場合、この選ばれた一つの候補字種 i を除いたほかの字種の中から評価関数 $d_S^2(\boldsymbol{x}, \boldsymbol{\mu})$ の値が最も小さい字種 j を選び (ただし、字種 jは候補とはしない)、式 (5.4) を用いて、評価を行う。$|z_{ij}(\boldsymbol{x})|/s_{ij} \leq \theta$ が満たされれば、字種 i を最終候補として出力する。満されない場合は適当な候補がないとしてリジェクトする。

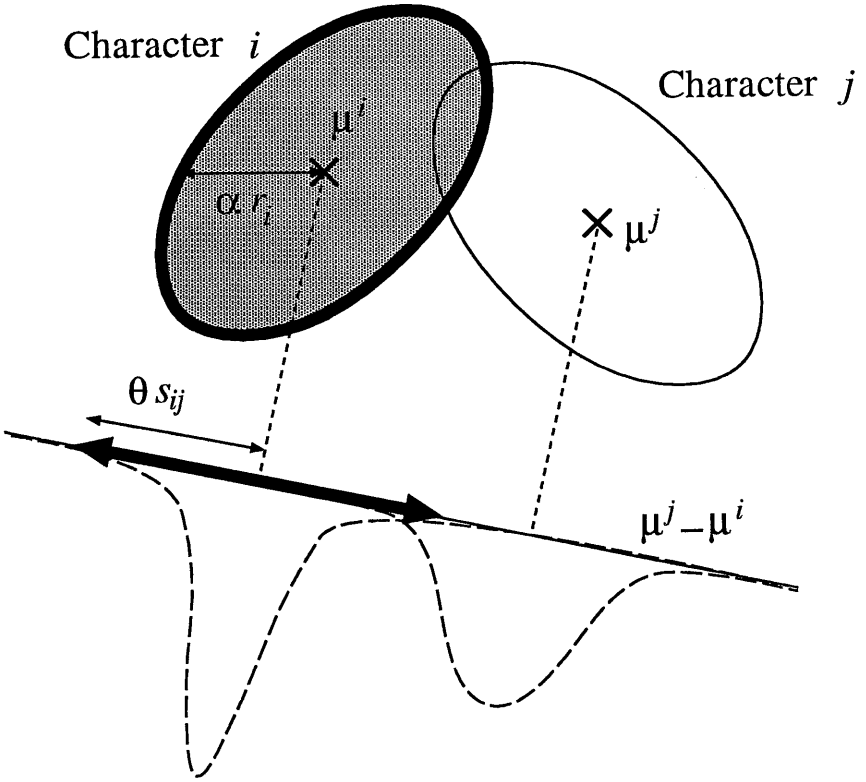


図 5.2: 特徴領域

なお、線形識別関数を用いず重心を結ぶ直線への写像を考えるのは、計算量削減とサンプル数が少ないことによる誤差の影響を軽減するためである。

式 (5.2) および式 (5.4) で表される特徴領域を図 5.2 中に太線で示す。太線の円は多次元の特徴領域を表し、太線の直線は一次元の特徴領域を表す。これらの範囲に含まれる字種を候補文字とすることになる。α と θ はパラメータであり、さまざまな値を取る。

5.4　簡素化マハラノビス距離

評価関数としてはよく整合しているほど値の小さい尺度であれば任意の尺度を用いることができるが、領域半径の概念を用いる場合にはサンプルパターンの分布をより正確に表せる評価関数が必要となる。評価関数としてマハラノビス距離を用いることが考えられる。マハラノビス距離は多次元正規分布の確率密度関数の式から導かれ、パターンが多次元正

規分布をしていると仮定でき、しかも十分なサンプル数を用いてパラメータを正確に推定できれば、パターンの分布形状の違いによらずあるパターンのあるカテゴリに所属する程度を正しく表すことができる。しかし、学習サンプル数が少ないと信頼性のある共分散行列を求めることができず、固有値展開した場合に特に高次成分で誤差が大きくなることが知られている [15]。また、共分散行列が正しく推定できたとしても、特徴量の次元数 Nに対して $O(N^2)$ の計算時間を必要とする等の問題がある。これらの問題を回避するためにマハラノビス距離と確率的に同等であり、少ない次元数 m $(m < n)$ で計算できる距離尺度として簡素化マハラノビス距離を提案し、用いることにする。サンプル数を十分に確保できない場合のこの距離尺度の振舞いについても実験的に明らかにする。

5.4.1　簡素化マハラノビス距離の定義

特徴量の次元数を Nとする。改良型方向線素特徴量の場合は $N = 196$ である。$\boldsymbol{x}$を未知入力ベクトル、$\boldsymbol{\mu}$、Σをそれぞれ平均ベクトル、共分散行列 $(N \times N)$ とすると、二乗マハラノビス距離 $d^2(\boldsymbol{x})$ は、

$$d^2(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{5.5}$$

となる (以下、二乗マハラノビス距離のことを単にマハラノビス距離とよぶ)。Σの第 k固有値をλ_k、λ_kに対応する固有ベクトルを$\boldsymbol{e}_k$とし、$\boldsymbol{y} = (y_1, y_2, ..., y_N)$ を $y_k = (\boldsymbol{x} - \boldsymbol{\mu}_k)^t \cdot \boldsymbol{e}_k$ で定義すれば、(5.5) 式は以下のように書き直せる。

$$d^2(\boldsymbol{x}) = \sum_{k=1}^{n} \frac{1}{\lambda_k} ((\boldsymbol{x} - \boldsymbol{\mu})^t \cdot \boldsymbol{e}_k)^2 \tag{5.6}$$

$$= \sum_{k=1}^{m} \frac{y_k^2}{\lambda_k} + \sum_{k=m+1}^{n} \frac{y_k^2}{\lambda_k} \tag{5.7}$$

ここで、式 (5.7) を次のように近似する。

$$d_S^2(\boldsymbol{x}) = \sum_{k=1}^{m} \frac{y_k^2}{\lambda_k} + \frac{1}{\lambda} \sum_{k=m+1}^{n} y_k^2 \tag{5.8}$$

$$= \sum_{k=1}^{m} \frac{y_k^2}{\lambda_k} + \frac{1}{\lambda} \left(\|\boldsymbol{x} - \boldsymbol{\mu}\|^2 - \sum_{k=1}^{m} y_k^2 \right). \tag{5.9}$$

$d_S^2(\boldsymbol{x})$ を 簡素化マハラノビス距離と定義する。式 (5.7) と (5.8) の第二項の期待値をそれぞれ E_1、E_2とする。λは定数であり、E_1と E_2が等しくなるように定める。式 (5.7) において、

λ_kはベクトルの $\boldsymbol{e}_k$ への射影成分の分散を表すから、パターンの分布を N次元正規分布と仮定すると $y_j/\sqrt{\lambda_j}$ は正規分布 $N(0,1)$ に従う。従って y_j^2/λ_j は自由度 1 のχ^2分布に従うため、式 (5.7) の第 2 項は自由度 $N-K$のχ^2分布に従い、$N-K$ が大きい (>100) 場合期待値 E_1 は

$$E_1 = n - m, \tag{5.10}$$

となる。一方、式 (5.8) において、$\boldsymbol{y}_j^2$ の期待値は$\boldsymbol{e}_j$ への射影成分の分散、すなわちλ_jに等しいため、期待値 E_2 は

$$E_2 = \frac{1}{\lambda}\sum_{j=K+1}^{N}\lambda_j \tag{5.11}$$

$$= \frac{1}{\lambda}\left(S - \sum_{j=1}^{K}\lambda_j\right) \tag{5.12}$$

となる。ただし S は Σ_i の対角成分の和である。$E_1 = E_2$とおけば、

$$\lambda = \frac{S - \sum_{k=1}^{m}\lambda_k}{n-m}. \tag{5.13}$$

が得られる。式 (5.13) を式 (5.9) に代入すると

$$d_S^2(\boldsymbol{x}) = \sum_{k=1}^{m}\frac{y_k^2}{\lambda_k} + \frac{1}{\lambda}\sum_{k=m+1}^{n} y_k^2 \tag{5.14}$$

$$= \sum_{k=1}^{m}\frac{y_k^2}{\lambda_k} + \frac{(n-m)(\|\boldsymbol{x}-\boldsymbol{\mu}\|^2 - \sum_{k=1}^{m} y_k^2)}{S - \sum_{k=1}^{m}\lambda_k}. \tag{5.15}$$

となり、平均ベクトル$\boldsymbol{\mu}$、Sおよび m 次元までの固有値・固有ベクトルのみを用いて $d_S^2(\boldsymbol{x})$ を求めることができる。

5.4.2　シミュレーション

提案した SMD が学習サンプルが少ない場合でも有効な距離尺度であることを示すため、SMD と真のマハラノビス距離との誤差についてシミュレーション実験を行い解析する。

まず適当な共分散行列Σと平均ベクトルμを与え[1]、この共分散行列を用い、乱数を用いて N次元正規分布 $N(\boldsymbol{\mu},\Sigma)$ に従うベクトルを n_t個発生させる。この n_t 個のベクトルから

[1]実際には ETL9B の最初の漢字である「亜」の字の方向線素特徴量を用いて計算した

共分散行列 $\hat{\Sigma}$ と平均ベクトル$\hat{\boldsymbol{\mu}}$ を求め、$\hat{\Sigma}$ から固有値$\hat{\lambda_k}$・固有ベクトル$\hat{\boldsymbol{e}}$ を求める。また、$\hat{\Sigma}$ の対角成分の和を $\hat{S}$ とする。そして、$\hat{\Sigma}$ を求めるのとは別にΣ から乱数を用いて n_e 個のベクトル (学習サンプル) を発生させ、それぞれのベクトルについて式 (5.9) により$\hat{\boldsymbol{\mu}}$, $\hat{\lambda_k}$, $\hat{\boldsymbol{e}_k}$, $(k = 1, ..., m)$ を用いて評価値 d_S^2 を計算する。Σ と$\boldsymbol{\mu}$ を用いて式 (5.5) により求めた評価値 d_{true}^2 を真のマハラノビス距離とし、d_S^2 と d_{true}^2 との誤差 e を $e = |d_S^2 - d_{true}^2/d_{true}^2|$ により計算し、n_e個のベクトルを用いた e の平均値を求める。

本研究と同様の考え方でマハラノビス距離の高次成分を用いず計算量を削減する評価値として、栗田らの擬似マハラノビス距離 (Quasi-Mahalanobis distance)[13], 孫らの改良型マハラノビス距離 (Modified Mahalanobis distance)[10] が提案されている。前者は式 (5.9) の λ として $m+1$ 次元目の固有値λ_{m+1}を用いるものであり (式 (5.16))、後者は式 (5.7) の第 2 項を無視しλ_jの代わりにバイアスを加えた$\lambda_j + b$ を用いるものである $(b = 5)$(式 (5.17))。

$$d_Q^2(\boldsymbol{x}) = \sum_{k=1}^{m} \frac{y_k^2}{\lambda_k} + \frac{n-m}{\lambda_{m+1}} \sum_{k=m+1}^{n} y_k^2 \tag{5.16}$$

$$d_M^2(\boldsymbol{x}) = \sum_{k=1}^{m} \frac{1}{\lambda_k + b} y_k^2 \tag{5.17}$$

しかし、これらの尺度はマハラノビス距離を近似することを目的として提案されたものではないため、真のマハラノビス距離との誤差についての解析はなされていない。これらの 2 つの尺度についても実験を行う。

$n_e = 10000$ とし、n_tの値を様々に変えてそれぞれの評価値を用いて誤差を求めた結果を図 5.3に示す。ただし、$m = 30$ とした。$n_e \geq 200$ のときは、$\hat{\Sigma}$ と$\hat{\boldsymbol{\mu}}$ を用いて式 (5.5) により求めた評価値 (Original Mahalanobis distance) を用いた場合の誤差も合わせて示してある (評価値の平均は Young の導出した理論式 [16] とよく一致していた)。

Original Mahalanobis distance を用いた場合、学習サンプル数 n_t が大きくなるに従って真のマハラノビス距離に限りなく近づくが、n_t が小さいときは他の距離尺度を用いた方が真のマハラノビス距離との誤差が小さいことが分かる。本研究で提案した尺度は $n_t = 100$ のときは孫らの尺度、栗田らの尺度よりも誤差が大きいが、$150 \leq n_t \leq 2000$ のときはいずれの距離尺度よりも誤差が小さくなっている。また、$n_2 > 2000$ のときも真のマハラノビス距離との誤差は 10%程度であり、提案した簡素化マハラノビス距離が真のマハラノビ

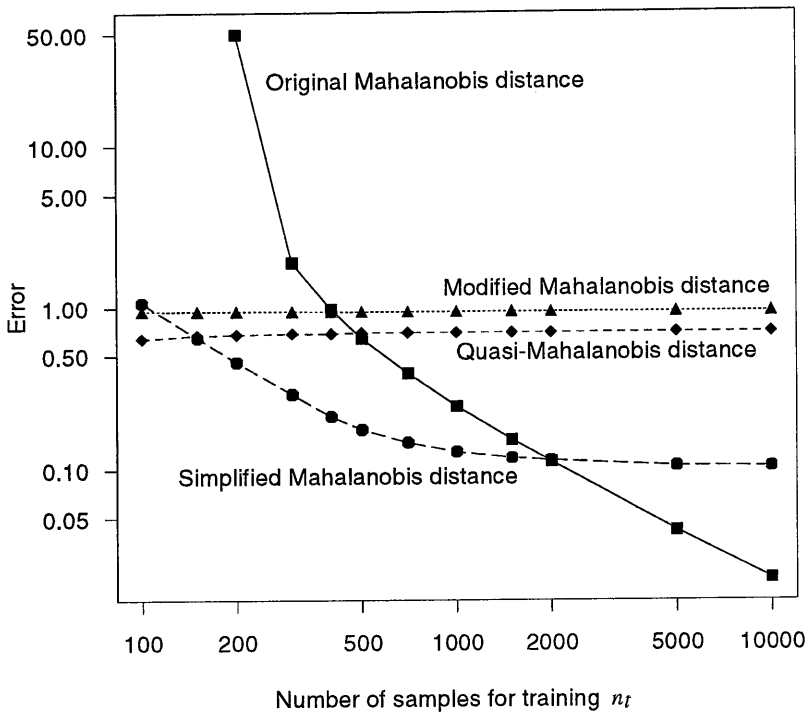


図 5.3: シミュレーションの結果

ス距離を近似する尺度として有効であることが確認できた。さらに、n_t が小さい場合は m を小さくすることで、n_tが大きい場合は m を大きくすることで誤差が減少する傾向にあることが確認された。これは、学習サンプルが少ないほど信頼できる固有値の数が減少することと、ここまで議論した近似の正当性を示すものである。

5.5　認識実験

5.5.1　認識性能の評価実験

本手法の有効性を確認するために認識実験を行った。ETL9B の 200 セットのデータの 20 セットずつを 1 つのグループとし、それぞれグループ A ～ グループ J と呼ぶ。ローテーション法を用い、1 つのグループを除いた 180 セットのデータを用いて認識用辞書を作成

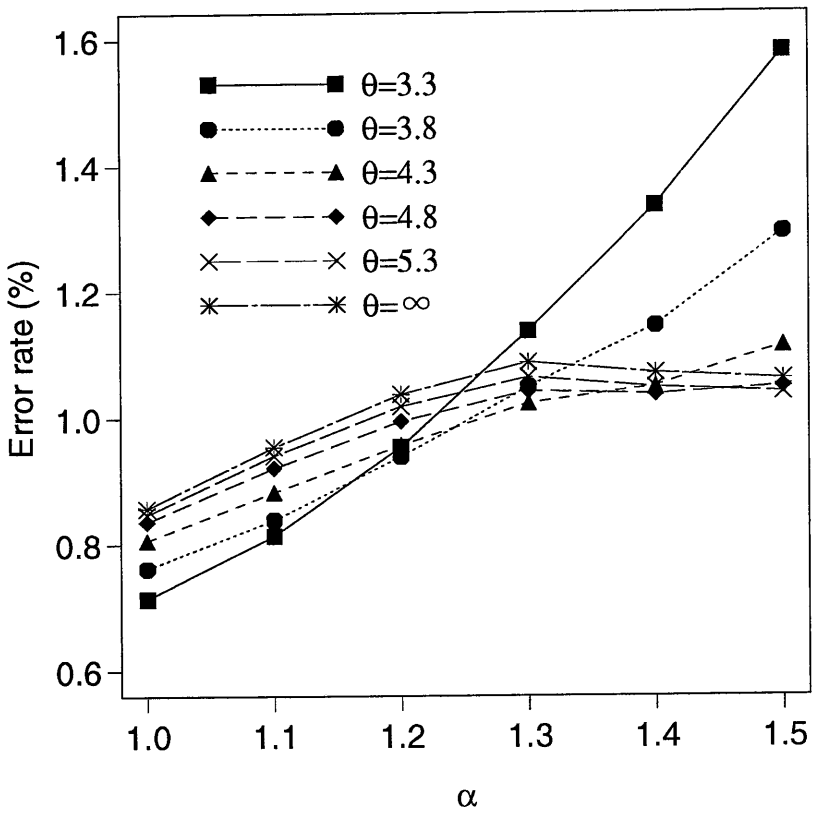


図 5.4: 誤り率

し、除いた 1 グループを評価用データとして認識する。この操作を各グループについて行い、誤り率 (候補群に正解が含まれない率)、平均候補数、リジェクト率を様々な α と θ を用いて求めた。10 グループの平均を取ったものをそれぞれ図 5.4、図 5.5 および図 5.6に示す。$\theta = \infty$ は、式 (5.2) の Cand($\boldsymbol{x}$) を最終候補集合とみなした場合の結果である。

α と θ の性質は次のようにまとめることができる。

- α は未知入力パターンから得られた特徴ベクトルと重心との距離の領域半径に対する比をどの程度まで許容するかを表す尺度である。
- θ はベクトルを一次元の特徴空間上に写像したときの許容範囲を表す値である。

図 5.4よりわかるように、誤り率を減少させるには α を小さくし、適当な θ を選べばよい。また、図 5.5より、リジェクト率は θ の値にかかわらずα が大きくなるに従って減少す

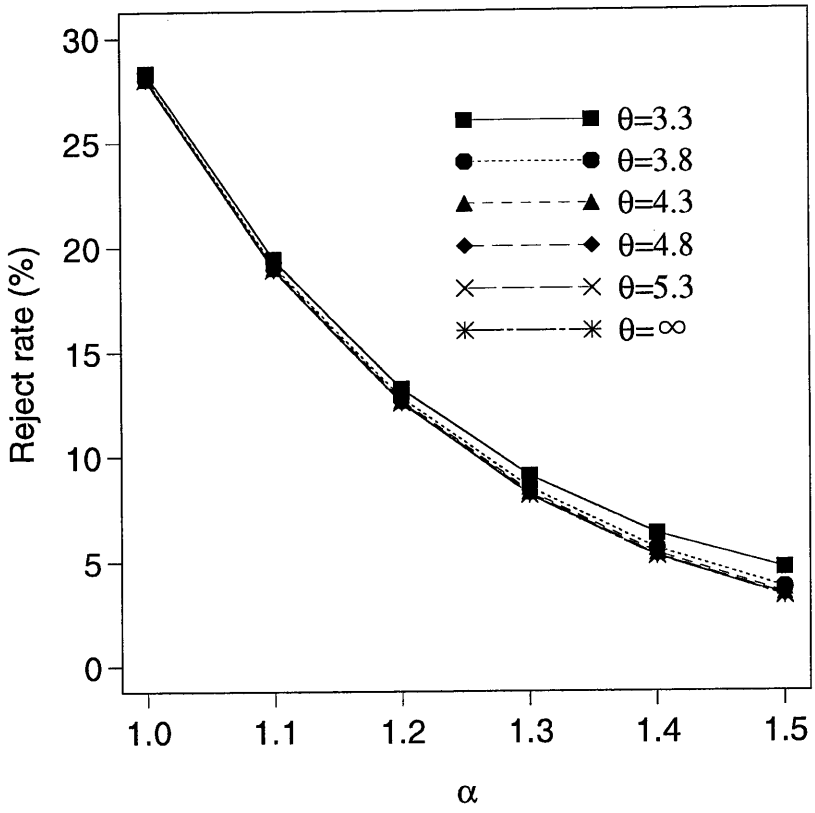


図 5.5: リジェクト率

る。これは、未知入力パターンがリジェクトされるのは式 (5.2) を満たす字種が存在しない場合がほとんどで、式 (5.4) によってリジェクトされることはあまりないことを意味する。しかし、これは一次元特徴領域が不必要であることを意味しない。図 5.6より、$\theta = \infty$ の場合平均候補数が急激に増加していることがわかる。一次元特徴領域を用いることで、候補を効率的に絞り込むことが可能になる。図 5.6は θ が小さいほど平均候補数も減少することを示している。すなわち、本手法は、式 (5.2) によって認識可能かどうかを判断し、式 (5.4) によって候補字種を限定するものであると言える。

以上より、本手法においては、単純に誤り率を小さくするには α を小さくして適当な θ の値を用いればよい。リジェクト率を下げるにはα を大きくすればよく、平均候補数を少なくするには α, θ ともに小さくすればよい。すなわち、適切なパラメータを選ぶことにより望ましい認識システムを構築することが可能になる。

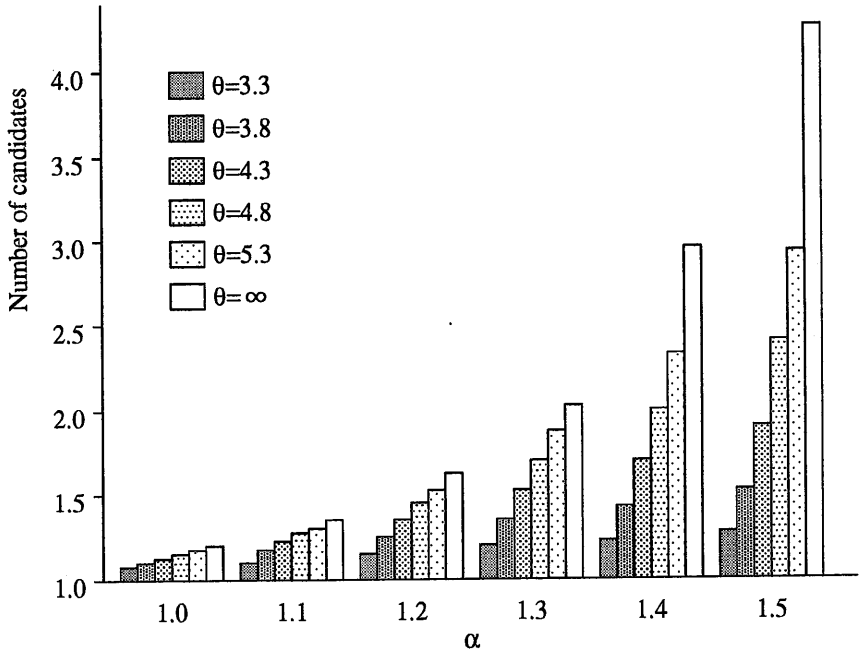


図 5.6: 平均候補数.

リジェクト率の許容範囲を約 10% とし、$\alpha = 1.2$ 、$\theta = 3.3$ の場合の各グループの認識率 (Rec.)、リジェクト率 (Rej.) および平均候補数 (Num.) を表 5.1に示す。また、比較のため、距離尺度としてユークリッド距離を用いて距離の最も小さいものを候補とする手法 (Euc.) についての結果も示してある。本手法により、ユークリッド距離を用いて必ず一つの候補を出力する手法と比較して誤り率を大幅に減少させることができる。たとえば、$\alpha = 1.2$ 、$\theta = 3.3$ の場合、誤り率は 0.95% となり、ユークリッド距離を距離尺度とした場合の 1/10 となった。これは、本手法が信頼性の高い認識候補の抽出に有効であると言える。さらに、この場合の平均候補数は 1.15 であり、多くの未知入力パターンにただ一つの認識候補しか選出してないことがわかった。また、提案するアルゴリズムに QMD を距離尺度として用いた場合の結果も表 5.1 に示した。提案した手法は比較すべき要素が複数あり (認識率・平均候補数・リジェクト率)、単純な比較はできないため、SMD のリジェクト率と同程度で認識率が最高になるようにパラメータを設定した ($\alpha = 0.88$、$\theta = 3.3$)。認識率と平均候補数を比較したところ、SMD の方が良い結果が得られ、提案した手法に適している距離尺度であることが確認された。これは、QMD と比べて、提案した SMD の方が特徴量の分布

表 5.1: Results.

Group	Proposed algorithm						Euc.
	SMD(α=1.2,θ=3.3)			QMD(α=0.88,θ=3.3)			
	Rec.	Rej.	Num.	Rec.	Rej.	Num.	
A	99.26%	9.04%	1.15	99.12%	9.35%	1.18	92.05%
B	99.01%	12.87%	1.15	98.68%	12.63%	1.18	90.27%
C	99.29%	13.30%	1.13	99.03%	13.41%	1.16	91.57%
D	99.08%	13.62%	1.15	98.78%	13.88%	1.18	90.03%
E	98.91%	13.61%	1.15	98.59%	13.84%	1.18	90.22%
F	99.02%	12.37%	1.17	98.63%	12.38%	1.20	89.19%
G	98.91%	14.98%	1.15	98.53%	15.00%	1.19	88.95%
H	99.07%	14.91%	1.15	98.63%	14.99%	1.17	90.01%
I	99.07%	11.22%	1.16	98.70%	11.19%	1.20	89.70%
J	98.83%	16.17%	1.15	98.38%	16.05%	1.18	89.12%
Ave.	99.05%	13.21%	1.15	98.71%	13.27%	1.18	90.11%

領域をより正確に表しているためであると思われる。

5.5.2　リジェクト性能の評価実験

1.1 で示したように、一般の文書認識システムにはリジェクト機能が不可欠であると思われる。ここで提案する手法のリジェクト性能を確認するための実験を行う。グループ A の 20 セットを評価用データとし、他の 180 セットを学習データにする。ETL9B の 3036 字種のうち 1 字種を除き、残りの 3035 字種の辞書を学習用の 180 セットのデータを用いて作成する。この辞書を用い、評価用の 20 セットに属し、辞書作成時に除かれた字種を認識する。3036 字種すべてに対して、同様な操作を繰り返す。$\alpha = 1.0$、$\theta = 3.3$ の場合、全字種のリジェクト率は平均で 90.28% となった。この結果より、辞書に含まれていない任意の未知入力パターンを認識する場合、高いリジェクト率が得られることが分かる。例えば切り出しのミスにより分離文字の一部を認識する場合等、実際に文字として存在しないパターンが入力された場合にこのリジェクト機能は大変有効であると思われる。

5.6　まとめ

本章では、文字パターンの多次元と一次元の 2 種類の特徴領域を統計的に推定した。多次元の特徴領域は、パラメータを正確に推定することができれば、文字認識に非常に有効である。一方、一次元の特徴領域は、識別能力は多次元の特徴領域と比べて低いが、サンプル数が少ないとき、一次元空間上のサンプルの分布を比較的正しく表すことができる。この二種類の特徴領域を効果的に用い、手書き文字を高精度に認識する手法を提案した。

多次元の特徴領域を表すために領域半径の概念を用いたので、評価関数として特徴ベクトルの分布を正確に表すものが要求される。そのための距離尺度として、マハラノビス距離と確率的に同等な尺度となる簡素化マハラノビス距離 (SMD) を定義し、これを用いた。SMD はサンプルが少ない場合において、マハラノビス距離と比べて高次成分の誤差による影響が少なく、計算時間も短縮でき、有効な距離尺度であると言える。

さらに、提案するアルゴリズムの有効性を ETL9B を用いた実験により確認した。許容リジェクト率を 10% 程度とすると、ユークリッド距離を用いて必ず 1 位候補を出力する場合と比較して SMD の方が誤り率を最大で 10 分の 1 に削減することができる。また、提案した認識手法で距離尺度として QMD を用い、SMD と比較したところ、SMD の方が良い結果が得られ、特徴領域を定義するのに、SMD の方が適した距離尺度であることがわかった。つまり、本研究で提案した SMD の方が、マハラノビス距離の統計的性質を保ったまま少ない次元数で特徴量の分布領域を表すことができると考えられる。

また、リジェクト機能も実験により示した。辞書に登録していない未知入力パターンを認識しようとするとき、リジェクトされる可能性が非常に高いことが分かった。

本手法は、文字の特徴領域を定義して未知パターンが属していると思われる字種をすべて候補とする点に特徴があり、最も確からしいと思われる候補を 1 個出力する従来の考え方と異なる。このような特徴は自然言語処理を用いる後処理の過程等において非常に役立つと思われる。なお、本手法は特徴量や距離尺度を制限するものではなく、一般的に用いることができる。本手法において信頼度を高くするとリジェクト率が高くなるが、リジェクトされたパターンは別の特徴量を用いて再認識する、という処理も可能である。これは今後の課題の一つである。また、ETL9B に限らず多様な手書き文字を用いて提案したアルゴリズムの有効性を確認し、さらに、後処理と組合せることでより実用的な手書き文字

認識システムを構築することも重要な課題である。

第 6 章

結論

本研究では、文字特徴量の分布形状を考慮した文字認識用辞書の構成法について検討を行った。ここで、活字と手書き文字のそれぞれの特徴量の分布形状の特徴に着目し、活字と手書き文字にそれぞれ適応と思われる二つの辞書構成法を提案した。

まず、特徴空間上の文字特徴量の分布形状を統計的な手法を用いて調査した。主成分分析を用いた調査から、同一字種の特徴量の分布の広がりに対して異なる字種の特徴量間の分布はかなり密であること、字種内の分布は特定の方向の分散のみが大きく、実際に領域が重なっている字種対は少ないこと、字種ごとに主成分の方向が大きく異なることがわかった。また、マルチフォント印刷文字と手書き文字の特徴量の分布形状の比較を行った。前者は分散は小さいが分布が単純な正規分布とは異なりピークが複数存在し、後者は分散は大きいが正規分布に近い分布となっていることが確認された。この調査結果を踏まえ、カテゴリ内分布とカテゴリ間分布の両方を考慮することにより誤認識の可能性のある字種のテンプレート数を増やす操作を繰り返すことで総テンプレート数を抑えたマルチテンプレートの辞書を作成する手法として、字種適応クラスタリング法を提案した。さらに手法の有効性について実験により確認した。単純なマルチテンプレート法で従来問題であった、不必要な複数化により認識用辞書が必要以上に大きくなり、記憶容量・計算時間が増加するなどの問題点を改善することができた。

手書き文字に対しては、特徴領域の推定による認識候補の高精度選出法を提案した。具体的には、まずサンプルパターンにより各字種の文字パターンの特徴量が分布する特徴領

域を統計的に定める。その際、一次元の特徴領域と多次元の特徴領域の二種類の特徴領域を用いた。これは、互いの利点を生かし、より信頼性の高い候補を得るためである。次に、領域半径の概念を用いて未知パターンが属している可能性のある特徴領域を推定し、候補選出を行う。さらに、得られた候補の信頼性を検証する。ここで、領域半径が用いられるため、パターンの分布をより正確に表す評価関数が必要となる。本研究では、確率的にマハラノビス距離と同等な距離尺度である、簡素化マハラノビス距離を提案した。これは少ない次元数で計算でき、従来のマハラノビス距離より計算時間が大幅に減少し、高次成分で生じる誤差の影響を軽減できる。シミュレーション実験により、学習サンプル数が少ない場合でもマハラノビス距離をよく近似し、有効な距離尺度であることを確認した。本手法は、文字の特徴領域を定義して未知パターンが存在していると思われる字種をすべて候補とするものであり、最も確からしいと思われる候補を 1 個出力する従来の考え方と異なる。さらに、リジェクト機能の導入より、手書き文字認識に適した柔軟性のある認識システムを構築することが可能になる。

これらの研究成果をもとに、確立された手法を文字認識に限らず一般的なパターン認識に応用することが今後の課題である。

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大学院工学研究科阿曽弘具教授、情報科学研究科丸岡 章教授、牧野正三教授に心より感謝致します。

また、情報処理教育センターの大町真一郎博士には本研究全般に渡り親身な御指導、御助言をいただきました。ここに深く感謝します。

さらに、御討論、御協力をいただいた根元研究室の加藤 寧博士、阿曽研究室の成富 敬博士、後藤英昭博士、森 大毅氏、鈴木基之氏に心から感謝します。

最後に多面に渡り御意見、御協力をいただき、また日頃の生活においてお世話になった阿曽研究室および丸岡研究室の皆様に感謝致します。

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), **J74-D-II**, No.3, pp.330–339 (1991-02).

[2] 阿曽弘具, 越後和徳, 木村正行 : “ 文字特徴量空間の性質と特徴抽出法の性能評価法 ”, 電子情報通信学会論文誌 (D-II), **J76-D-II**, No.11, pp.2285–2294 (1993-11).

[3] 目黒眞一, 梅田三千雄 : “ マルチフォント印刷漢字の認識 ”, 電子情報通信学会論文誌 (D), **J65-D**, No.8, pp.1026–1033 (1982-08).

[4] 郭 軍, 孫 寧, 根本義章, 佐藤利三郎 : “ 整形変換を用いた手書き文字データベース ETL9B の高精度認識 ”, 電子情報通信学会論文誌 (D-II), **J76-D-II**, No.5, pp.1015–1022 (1993-05).

[5] 泉井良夫, 原島 博, 宮川 洋 : “ 階層的な辞書の変形を用いた手書き文字認識 ”, 電子情報通信学会論文誌 (D), **J68-D**, No.3, pp.361–368 (1985-03).

[6] 塩野 充 : “ 多重類似度法による手書き漢字識別の基礎実験 ”, 情報処理学会論文誌, **27**, No.9, pp.853–859 (1986-09).

[7] 八代博昭 : “ 手書き漢字認識の高精度化に関する基礎研究 ”, 東北大学大学院工学研究科情報工学学専攻 昭和 63 年 修士学位論文 (1988-02).

[8] 加藤 真, 高橋弘晏 : “ 階層的辞書配置によるマルチフォント漢字認識 ”, 電子情報通信学会論文誌 (D-II), **J74-D-II**, No.1, pp.8–18 (1991-01).

[9] 斉藤泰一, 山田博三, 山本和彦 : "JIS 第 1 水準手書漢字データベース ETL9 とその解析", 電子情報通信学会論文誌 (D), J68-D, 4, pp.757–764 (1985-04).

[10] 孫 寧, 安倍正人, 根元義章 : "改良型マハラノビス距離を用いた高精度な手書き文字認識", 電子情報通信学会技術研究報告, PRU94-94, pp.65–72 (1994-11).

[11] 若林哲史, Yang Deng, 鶴岡信治, 木村文隆, 三宅康二 : "非線形正規化と特徴量の圧縮による手書き漢字認識の高精度化", 電子情報通信学会技術研究報告, PRU95-1, pp.1–8 (1995-05).

[12] 孫 寧, 安倍正人, 根元義章 : "改良型方向線素特徴量および部分空間法を用いた高精度な手書き文字認識システム", 電子情報通信学会論文誌 (D-II), Vol.J78-D-II, No.6, pp.922–930 (1995-06).

[13] 栗田昌徳, 鶴岡信治, 横井茂樹, 三宅康三 : " 加重方向ヒストグラムと疑似マハラノビス距離を用いた手書き漢字・ひらがな認識 ", 電子情報通信学会技術研究報告, PRL82-79, pp.105–112 (1983-01).

[14] Y.Linde, A.Buzo, and R.M.Gray, "An Algorithm for Vector Quantizer Design", *IEEE Trans. on Comm.*, Vol. COM-28, No.1, pp.84–95, Jan 1980.

[15] 竹下鉄夫, 木村文隆, 三宅康三 : " マハラノビス距離の推定誤差に関する考察 ", 電子情報通信学会論文誌 (D), **J70-D**, No.3, pp.567–573, (1987-03).

[16] Young, I.T., "Further Consideration of Sample and Feature Size," *IEEE Trans. Inf. Theory*, vol.IT-24, no.6, pp.773–775, June 1978.

業績一覧

[1] 孫 方, 大町真一郎, 阿曽弘具 : "カテゴリー間分布を考慮した文字認識用辞書のマルチテンプレート化の一手法", 電子情報通信学会技術研究報告, PRU94-11 (1994-05).

[2] Fang Sun, Shin'ichiro Omachi *and* Hirotomo Aso, "Precise Selection of Candidates for Handwritten Character Recognition Using Feature Regions," IEICE Trans. Inf. & Syst. (条件付き採録).