氏名	孫（そん） 方（ほう）
授与学位	博士（工学）
学位授与年月日	平成１１年３月２５日
学位授与の根拠法規	学位規則第４条第１項
研究科、専攻の名称	東北大学大学院工学研究科（博士課程）電気・通信工学専攻
学位論文題目	特徴量の分布形状を考慮した高精度文字認識に関する研究
指導教官	東北大学教授 阿曽 弘具
論文審査委員	主査 東北大学教授 阿曽 弘具　東北大学教授 矢野 雅文 東北大学教授 丸岡 章

論文内容要旨

1 序論

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．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．このような計算機の発展にともなって，新聞，雑誌，書籍などの文書の管理を計算機によって行おうという社会的要請が強まってきた．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ますます重みが増していくと考えられる．

本論文では，印刷文字及び手書き文字の認識の高速化・高精度化を目的とし，文字特徴量の分布形状を考慮した文字認識手法を開発，提案している．

2 文字認識

一般的な文書認識は画像入力，文字切り出し，特徴抽出，候補選出，結果出力の５つの処理からなり，各分野の成果を統合してはじめて１つの文書認識システムが完成する．このうち，個々の文字パターンに対してそのパターンが表す文字が何であるかをいかに精度良く判別するかが狭義の「文字認識」と呼ばれる分野に属する．

文字認識の性能を左右する重要な要素として，辞書と識別関数が挙げられる．辞書とは認識対象となる各字種のパターンまたはパターンから抽出された特徴量の集合である．識別関数とは，入力パターンと辞書パターンの類似性を評価するための尺度である．本研究では，これらを重点的に検討した．認識のための特徴量としては改良型方向線素特徴量を用いた．

3 カテゴリ間の分布を考慮したマルチテンプレート辞書の構成法

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，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した．具体的には，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) を提案した．

JIS 第一水準の数字・アルファベット・平仮名・片仮名・漢字 (計 3163 字種) に対する本手法の有効性を確認する実験を行った．実験データは，明朝体 12 種類，ゴシック体 12 種類の計 24 種類のフォントの文字である．認識手法は本手法，従来法 (1 字種あたりのテンプレート数が 4 の LBG 法) およびシングルテンプレート法である．シングルテンプレート法の認識率は 97.61%であるのに対し，従来法と本手法の認識率はともに 98.46% となり，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，計算時間・記憶容量ともに削減できた．これにより本手法の有効性が確認された．

4 特徴領域の推定による高精度候補選出法

本章では，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した．本アルゴリズムの特徴は，候補選出の際に，多次元と一次元の 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．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，多次元特徴領域と一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，信頼性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) を提案した．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，識別に有効な距離尺度であると言える．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した．

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した．その結果，従来より提案されていた距離尺度である擬似マハラノビス距離 (QMD) よりも SMD の方が良い結果が得られ，特徴領域を定義するには SMD の方が適した距離尺度であることがわかった．これは，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たためと考えられる．

5 高次元ベクトルの混合分布推定アルゴリズム

統計的パターン認識手法において，高い認識精度を達成するためにはパターンの分布形状を正確に推定することが重要である．認識対象から抽出した特徴ベクトルの分布形状は複雑で多様であるにもかかわらず，多くの場合において，簡単のため分布が正規分布であると仮定される．実際には，このような複雑な分布を表わすためには混合正規分布のほうが望ましいと思われる．しかし，文字認識のように特徴量が高次元のパターン認識問題では，混合分布のパラメータを正確に推定するには大量の学習サンプルを必要と

するため，学習サンプルが十分に得られない状況下で混合正規分布が用いられることは極めて少なかった．本章では，最尤推定に基づいて混合正規分布のパラメータを推定するための新しいアルゴリズムを提案した．簡素化マハラノビス距離をこの推定過程に導入することによって，学習サンプルが十分に得られない場合においても，パラメータの推定を実現することが可能となる．

また，提案したアルゴリズムを文字認識用のマルチテンプレート辞書の構築に適用した．辞書は，クラスタリングとクラスタ間情報を利用した混合分布推定を繰り返すことによって構築される．本手法を用いることで，分割する必要のあるカテゴリのみ分割することが可能となる．また，一つのカテゴリをいくつのクラスに分割しても，各クラスのパラメータを推定する際に，そのカテゴリに属するすべての学習サンプルを使うので，分割することによるパラメータの信頼性の低下が少ない．

本手法の有効性を手書き文字を用いた認識実験より確認した．認識対象として，ETL9B の中で誤認識の可能性の高い字種である仮名の 71 字種を用いた．辞書として，本手法で作成した辞書のほか，分割を行わない辞書 (single-template 辞書と呼ぶ)，すべてのカテゴリを 2 分割した辞書 (two-class 辞書と呼ぶ) を用いた．3 種類の辞書を用いた認識結果より，提案する辞書を用いた場合の誤り率が最も小さいことがわかった．この結果から，提案した混合分布推定アルゴリズムが高次元特徴ベクトルを用いた認識問題に対して有効であることが確認された．本手法によって two-class 辞書よりも良い結果が得られたことは，必要なものだけを分割することで，誤認識の可能性を減少させ，同時に計算コストの削減を実現することが可能であることを示すものである．

6 特徴量の要素の相関を考慮した高速・高精度な識別関数

近年提案された手書き文字認識手法の多くは高い認識率を得るために膨大な計算量を必要とする．高精度な認識システムを構築することができた今，認識性能を落とさずにいかに計算量を削減するかは文字認識における重要な課題の 1 つである．

本章では，ベクトルの各要素間の相関に着目してマハラノビス距離の近似値を計算できるベクトル分割型準マハラノビス距離を提案し，文字認識に応用した．これは特徴ベクトルを分割してマハラノビス距離を計算する手法であり，高い認識性能を保ちながら計算量を削減できる．また，学習に用いるデータ数に制限がある場合も相対的に次元数に対する学習データ数を増やすことが可能になり，データ不足の問題が改善できる．まず文字認識における特徴量の共分散行列を求め，特徴量の各要素の相関を分析し，ベクトルを分割することでマハラノビス距離を近似できることをシミュレーション実験で検証した．さらに，共分散行列をより分割に適した形に変形するためのアルゴリズムを提案し，その効果を確認した．最後に，ETL9B を用いた認識実験を行い，本手法が文字認識に有効であることを確認した．

提案手法を従来法 (改良型マハラノビス距離:MMD) と比較した結果，従来法より高い 99.06%の認識率が得られることが分かった．つまり，本手法では共分散行列の対角付近から遠い成分を 0 で近似しているものの，その影響は従来法で行われてきた単純に高次成分を無視することによる影響よりもはるかに少ないと言える．また，本手法により計算時間を従来法の約 1/2 に削減できることも確認できた．これらの実験結果から，ベクトル分割型準マハラノビス距離は，データが少ない場合にもより信頼性があり，認識性能を落とさずに計算量を大幅に削減できる高速・高精度な識別関数であることが分かった．

7 結論

本研究では，文字の特徴量の空間上での分布形状を考慮し，マルチテンプレート辞書作成法と識別関数に重点を置いて統計的パターン認識手法を検討し，高速・高精度な文字認識手法を実現した．

審 査 結 果 の 要 旨

人間とコンピュータの間の柔軟なマン・マシンインターフェースを実現するためには、文字認識の高精度化・高速化を図る技術が不可欠である。一般に高精度なパターン認識を実現するためには、認識対象パターンが分布する領域の形状を推定し、それに適した認識方法を用いる必要がある。しかし文字認識の場合、高次元の特徴量を必要とすることから分布形状を把握することが難しく、これまで認識に際し分布形状を考慮することはほとんどなかった。著者は、文字特徴量(特徴ベクトル)の分布形状を考慮した文字認識手法および識別関数について研究を行い、文字認識の高速化、高精度化を実現した。本論文はこれらの成果をとりまとめたもので、全編7章からなる。

第1章は序論である。

第2章では文書認識処理の流れを示すとともに、本研究で用いた文字特徴量について述べている。

第3章では、マルチフォント印刷文字の認識手法として字種適応クラスタリング法を提案している。この手法は、特徴ベクトルの空間上での分布の様子から誤認識の可能性のある字種を求め、字種ごとに適切な数のテンプレート(辞書ベクトル)を組織的に構成するもので、字種によらず固定した数のテンプレートを用意する従来手法に比較して、認識率、使用メモリ量の点で優れていることを認識実験により明らかにしている。これは有用な成果である。

第4章では、文字認識のための識別関数として用いるマハラノビス距離を少ないサンプルで適切に近似できる簡素化マハラノビス距離を提案し、それを用いて文字特徴量の分布領域を求め、その領域に属するか否かを判定することにより認識する新しい認識手法を提案し、手書き文字を用いた実験によりその有効性を明らかにしている。簡素化マハラノビス距離の近似性能を実験的に評価するとともに、少ないサンプルで構成できる既存の識別関数に比較して近似性能だけでなく認識性能でも優れていることを実証している。これは実用上有用な知見である。

第5章では、特徴量の分布形状を正確に表す混合正規分布のパラメータを推定する新しい手法を提案し、これに基づいた認識手法を与えている。この推定法は、簡素化マハラノビス距離の特性を利用して、従来推定が不可能とされていた学習サンプル数が極端に少ない場合の推定を可能にしたもので、重要な成果である。また、提案した認識手法は、誤認識の可能性が高い字種に限り混合正規分布をもとにした識別関数を用いるように改良したマルチテンプレート法であり、手書き文字を用いた認識実験により有効性を検証している。

第6章では、特徴ベクトルに関する共分散行列の特性に着目し、ベクトルの要素を分割してマハラノビス距離の近似値を計算するベクトル分割型準マハラノビス距離を提案し、その妥当性を実験により検証している。また、この識別関数を用いる際のベクトルの適切な分割法を与え、手書き文字認識実験で、少ない学習サンプルで従来法の認識性能を保ったまま処理速度において従来法の約1/2となることを明らかにしている。

第7章は結論である。

以上要するに本論文は、文字特徴量の分布の様子を考慮したマルチテンプレートの作成に基づく認識手法を与えるとともに、特徴量の分布形状を考慮することにより高精度な識別を可能とする識別関数を考案し、その有効性を実証したもので、パターン認識工学及び情報通信工学の発展に寄与するところが少なくない。

よって、本論文は博士(工学)の学位論文として合格と認める。