修士学位論文

論文題目　文字パターンに基づく
文字の符号化に関する基礎的研究

提出者　東北大学大学院工学研究科
電気及び通信工学　専攻
学籍番号　5m167
氏名　山田光彰

指導教官	阿曽 弘具 教授
審査委員 （○印は主査）	○ 阿曽弘具 教授 1 丸岡章 教授 2 瀬尾澤正道 教授 3 教授 4 教授

提出者略歴	
氏名（ふりがな） やまだ みつかげ 山田 光影	昭和 46 年 2 月 12 日生
本籍 栃木 都道府県	
履歴事項	
~~昭和~~ 平成 元 年 4 月 1 日	東北 大学 工 学部 学科入学
~~昭和~~ 平成 5 年 3 月 25 日	同 卒業
~~昭和~~ 平成 5 年 4 月 1 日	東北大学大学院工学研究科 電気及び通信工学 専攻 前期2年の課程入学
昭和 平成 年 月 日	同 修了
昭和 平成 年 月 日	
昭和 平成 年 月 日	
昭和 平成 年 月 日	
昭和 平成 年 月 日	

備考(1) 履歴事項は、大学入学から年次にしたがって記入すること。

(2) 修士課程の修了年月日は、学位記授与式年月日を記入すること。

修士学位論文

文字パターンに基づく文字の符号化に関する基礎的研究

東北大学大学院工学研究科電気及び通信工学専攻

山田 光影

目 次

図目次

表 目 次

第1章

序論

1.1 本研究の背景

電子計算機は今世紀最大の発明品の一つとして知られ，情報産業の中核的存在となったばかりでなく，情報化社会の到来に対して先導的役割を果たし続けてきた．電子計算機がこの世に現れたとき，人々はその斬新な機能の未来に大きな夢を託して，これに「人工知能」という異名を贈った．

しかし，今日の電子計算機の使われ方を見ると，それよりはむしろ万能情報処理機械と呼んだほうが，その名にふさわしいように思われる．電子計算機が最も得意としている情報処理作業がもちろん数値計算であることに異論はないだろう．ところが，電子計算機の究極の目標として当初から掲げられていた人間の頭脳にとって，計算はあまり得意な部類に属する情報処理作業ではなかった．

パターン認識は，人間は勿論あらゆる動物達の頭脳にとって，最も得意な情報処理作業であることが知られている．しかも，皮肉なことに，このパターン認識は，電子計算機が長年にわたって最も苦手としてきた情報処理作業の代表例であった．人工知能の異名で呼ばれてきた電子計算機と人間本来の頭脳との間にある機能上のこの相違は，宿命的なものであるとして，長い間懸案の課題と目されてきた．

パターン認識の一分野である文字認識は30年以上研究され続けてきた．研究が始められた当時の計算機の能力は今とは較べものにならなかったほど低かったので，とても実用に供し得るようなパターン認識装置は開発できる筈がなかった．十数年に及ぶ基礎研究が積まれた後，文字読み取り装置 (OCR:Optical Character Reader) の開発が軌道に乗ったのは，昭和40年代に入ってからである．郵便番号の自動読み取り区分機の開発が成功して，郵便番号制が実施されたのは丁度その頃である．当時の技術水準は，自由手書き文字の読み取りは数字十種類程度が限度で，英数字・片仮名等の文字読み取りは印刷文字に限られていた．しかし，その後昭和50年代に入った頃，LSI技術を基盤とするマイクロプロセッサの発明がなされ，これによってOCRの著しい低廉化が図られたため，急速に普及が進行するようになり，遂に英数字はもとより数千種にのぼる手

書き漢字さえも読み取りが可能な段階にまで到達したのである[1].

OCRにより既存の印刷物を電子計算機で取り扱えるデータに変換することにより，印刷物のデータベース化に携わる人間の労力を削減できる．また，活字本を自動的に点字本に翻訳する自動点訳装置により，既存の本の点訳がより簡単になる．さらに，ペンタッチによりオンラインで文字を認識し文字入力が可能な電子手帳も既に商品化されている．このように，様々な分野での応用が期待される文字認識技術は，今後もその重要さが増すと考えられる．

現在，個別文字認識に最も求められているものが二つある．一つは認識速度でありもう一つは認識精度である．

認識速度について，現在一般に用いられている全数整合法を例に挙げて考えてみる．全数整合法では，一つの文字にあらかじめ一つの標準パターンが用意されており，ある未知入力文字パターンは全字種の標準パターンと比較され，最も似ている標準パターンを認識結果としている．したがって比較回数は全字種数と等しく，字種数の多い日本語の文字認識では，この比較回数の多さが計算時間を増大させる原因となっている．

従来は，比較回数を低減するために，大分類という手法がとられてきた．この手法の基本的な考え方は，全字種を類似する字種のクラスに分割し，まず，未知入力文字がどのクラスに属するかを決め(候補の絞り込み)，次いで，そのクラス内のどの字種であるかを決めるということである．これにより比較回数は(クラス数+クラス内文字数)となり，全数整合法に較べ比較回数を低減できる．この大分類法での問題点の一つは，未知入力文字のクラス分けで間違うと，誤りを回避できないことである．しかも，クラス分けを間違わないようにするためのクラス分割の方法に確実なものがなかったことである．

いっぽう，認識精度について，漢字を例に挙げて考えてみる．漢字には「体」「休」「保」のように類似文字が存在する．そして，これらの類似文字は局所的にのみ異なっている．したがって，全体のパターンを比較すると，パターン間の差が小さくなり識別が困難となる．そこで，局所的に異なる部分だけのパターンを比較することで，識別の精度を上げることが期待できる．また，多くの漢字は部首(偏，冠，傍など)を含んでいる．漢字には部首だけが同じというものがあるので，これらの部首の抽出ができると，残った部分の識別だけを行なえばよく，認識は比較的簡単になる．これは部首によるクラス分けに基づく大分類法であり，比較回数の削減という利点もある．

分類を効率よく行うという観点から漢字の部首抽出が試みられているが[2][3]，高精度に部首を同定することは困難であることが知られている[4]．また，部首以外の部分情報を利用する研究も行なわれている．孫らは，類似文字の識別に必要な部分情報を自動学習によって獲得し，詳細識別に利用する認識手法を提案した[5]．しかし，この手法では，文字の構造に基づいてあらかじめ定めた部分について考えており，その大きや形についての妥当性は示されていない．江島らは，特徴ベクトルを部分ベクトルに分割し，個々

の部分ベクトルで得られた分類情報を再度全体にわたって統合することにより，手書き文字の大分類を行う手法を提案した[4]．この手法では，マスクを用いて文字パターンから部分ベクトルを抽出しており，マスクの大きさに関する検討はなされているが，文字の構造に関しては言及されていない．したがって，あらかじめ定めずに，文字の構造に関する部分の大きさや形を知る必要がある．

1.2 本研究の目的

文字パターンは一定数の部品の組合せから成ると考えることができ，各部品の種類と部品の数の組合せから作られるパターンが字種数以上であれば，全ての文字パターンを表現できる．またこのとき，部品の各種類に一つの記号を割り当てれば，文字パターンは，部品の種数と等しい種数の記号を持ち，部品の数と等しい長さの記号列に変換できる．つまりこれは，文字パターンに基づく符号化を意味する．これが実現できると，(部品の種数×部品の数) だけの比較回数で文字を決定することが可能となる．これにより，全数整合法に較べ比較回数を減らすことができる．つまり，文字パターンを適当な部品に分割することができれば，少ない比較回数で文字の決定が可能となる．これは，一つの部品における比較を大分類と同様なものとみなせることを意味する．さらに，類似文字どうしでは，組合された部品のうち異なると思われる部品だけを比較することで詳細な識別が可能となる．しかし，文字パターンの部品を部首などと考えると，その切り出しが困難であることが知られている．そこで，文字パターンの領域全体の一部である部分領域を考え，各部分領域での文字パターンを部品とする可能性を考える．このとき，

- 部分領域の選び方
- 部分領域の中でパターンをいかに，かつ，いくつに分類するか

が問題となる．本研究では，認識に有効な符号化法を考案することを目標に，上の二つの項目について検討し，文字パターンの部品構成を組織的に求めるための基準を明らかにし，これを基にした一つの符号化法を提案し，その性能評価を行うことを目的とする．

1.3 本論文の構成

第 1章 序論であり，本研究の背景および目的を述べる．

第 2章 本研究で特徴量として用いた方向線素特徴量の抽出について述べる．

第 3章 部分領域の安定度を導入し，部分領域における文字パターンの性質を調査する．

第4章 第3章の結果から符号化に有効な部分領域の組合せを作り，それを基にした文字パターンの符号化による大分類について検討する．

第5章 結論であり，本研究のまとめおよび課題を述べる．

第 2 章

方向線素特徴量の抽出

文字イメージは多次元のパターンベクトルで表わされるが，パターンベクトル空間の次元数はかなり大きいので，このままでは処理に費やされる計算量が膨大になり好ましくない．したがって，文字イメージは，前処理および特徴抽出の段階を経て，より少数の数値の組または記号の組に変換される必要がある．つまり，特徴抽出の段階で要求される処理は，パターンベクトル空間でのパターンベクトルの分布状態をなるべく損なわないようにしながら，その次元数を極力減らすことであると言える．この特徴抽出の段階を経て得られる数値の組または記号の組は，イメージの本質的で有効な情報を含むものであり，特徴量と呼ばれる．一般に特徴量が数値の組の場合は特徴ベクトルとも呼ばれる．

本研究では特徴量として方向線素特徴量[6]を用いる．この特徴量は，比較的次元数が小さく，文字の局所的な構造を反映しているという特徴を持つ．本研究では文字の部分情報を利用し，より高速な認識手法を提案するので，文字の局所的な構造を反映する方向線素特徴量を用いることは妥当であると考える．

そこで本章では，文字イメージから方向線素特徴量を得るまでの処理を述べる．処理の流れは図 2.1の様になっている．以下に各処理過程について詳しく述べる．

2.1　文字入力

認識対象となる文書は光学式スキャナにより二値化され入力される．その後，入力されたイメージから各文字イメージの領域を抽出する処理である文字切り出しを行う．横書きの文書の場合，文字は列方向には揃っていないが行方向に揃っていることが多く，また，一般に文字の縦方向の文字間隔は横方向の文字間隔より大きいので，まず全体のイメージデータから行を切り出す．次に，この一行分の画像から一文字ごとの領域を切り出す．こうして 1 文字ごとに切り出されたイメージを次の処理への入力とする．

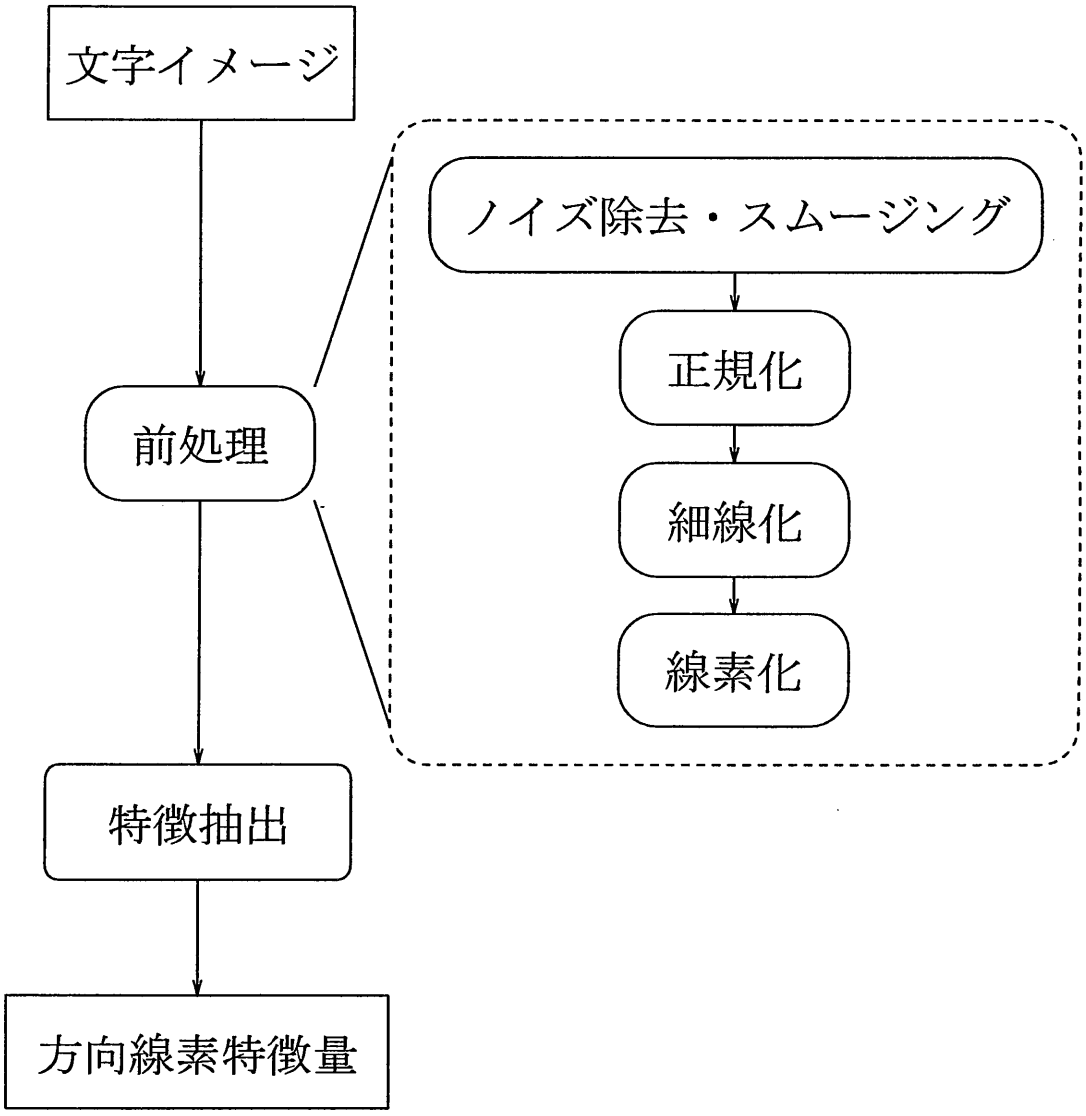


図 2.1　特徴抽出の流れ

2.2　前処理

前処理とは後段の処理をしやすくするために行われる予備的な処理である．入力された文字の特徴量を求めるために，切り出された文字のイメージに対して以下の四つの前処理を行う．

2.2.1　ノイズ除去・スムージング

入力された二値画像イメージには，印刷時のインクリボンの状態やスキャナの性質などによって，ノイズや線分上の凹凸が少なからず発生している．これらは認識精度に悪い影響を及ぼすため除去される必要がある．まず，図 2.3(a) にあるような 2 ドット×2 ドット以下の孤立点をノイズとみなし，これを除去する．次に 3 ドット×3 ドットのマスクを用いてスムージング (線分の円滑化) を行う．図 2.2(a) は凸となっている 3 ドット×3 ドットのマスクの一番上を除去する例であり，図 2.2(b) は凹となっている 3 ドット×3 ドットのマスクの中心を埋める例である．ノイズ除去，スムージング後のイメージを図 2.3(b) に示す．

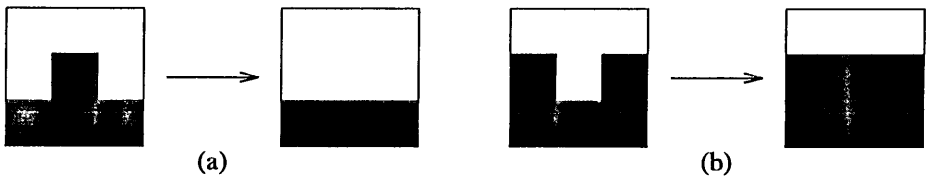


図 2.2　スムージングのためのマスク

2.2.2　正規化

印刷文字の大きさは一般に 4 倍角・縦倍角・横倍角・全角など様々である．よって，切り出された文字の大きさや位置の違いによる影響を吸収するために，元のイメージを一定の大きさ (本研究では 64 ドット×64 ドット) に拡大または縮小する必要がある．これを正規化と呼ぶ．最も単純な正規化は，縦・横両方向に一定の大きさにイメージを線形伸縮することで実現できる．これを線型正規化と呼び，本研究でもこの線型正規化を行うものとする．正規化後のイメージを図 2.3(c) に示す．

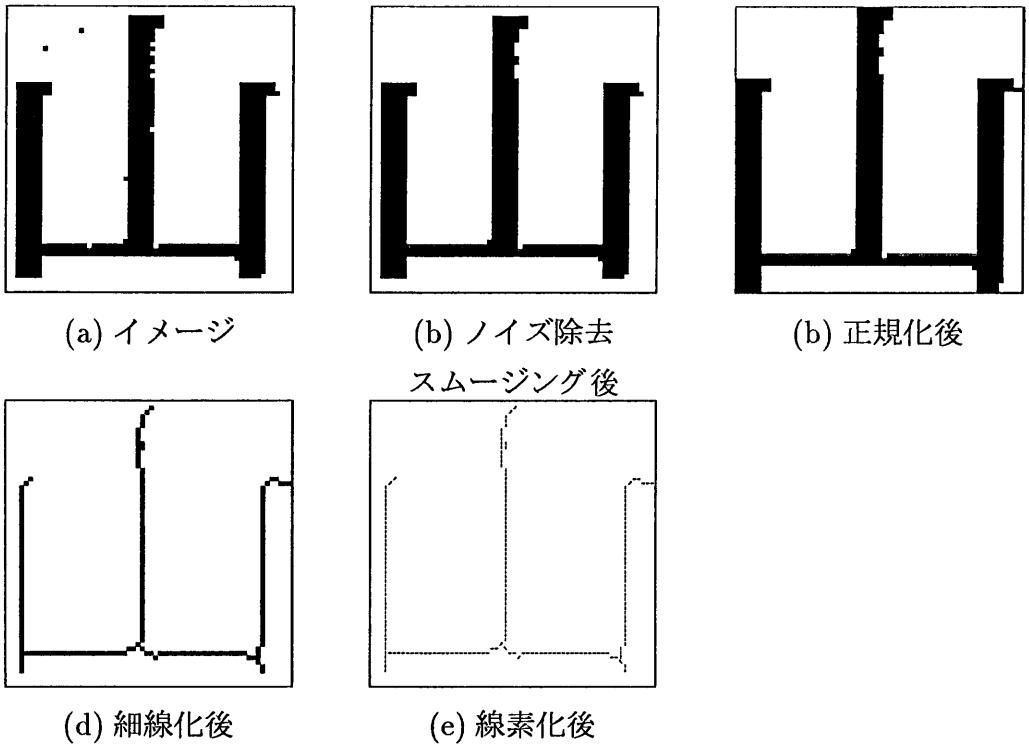
2.2.3　細線化

正規化後の文字イメージは，同じ字種であっても異なるイメージでは文字線幅が異なることがある．そこで，文字線幅の影響を吸収するために細線化を行う．細線化はイメー

ジにおいて幅が 2 ドット以上の線分を幅が 1 ドットの線分に変換する処理である．本研究では Hilditch の方法[7]を用いる．この方法は，各画素においてその 8 近傍に注目し，8 連結性をもとにイメージを 1 ドットずつ削っていく方法である．イメージの変化がなくなるまでこれを繰り返す．即ち，元のイメージの線分の幅の半分の回数を繰り返すことにより，最終的に線分の幅が 1 ドットの連結図形が得られる．ところが，潰れた文字イメージに対して線分の幅を 1 ドットにまで細線化すると，潰れの部分で文字が本来持っている情報が失われてしまう．このため，細線化回数に上限を設け，その細線化回数で幅が 1 ドットにならない部分に対しては，その輪郭線を求めることとする．この処理を行うとき，細線化回数の上限は文献 [6] で 5 回が最適とされており，本研究でもこの値を用いる．細線化後のイメージを図 2.3(d) に示す．

2.2.4　線素化

細線化されたイメージの微少領域の方向を規定し，局所的な特徴を得るために線素化を行う．細線化されたイメージの各黒画素について，その画素を中心とする 3 ドット×3 ドットの領域を参照し，縦 (｜)，横 (—)，斜め 45 度 (／)，斜め 135 度 (＼) の 4 方向の線素のうち，最も自然であると考えられる一つの線素をその画素に対応させる．これを線素化と呼ぶ．線素化後のイメージを図 2.3(e) に示す．



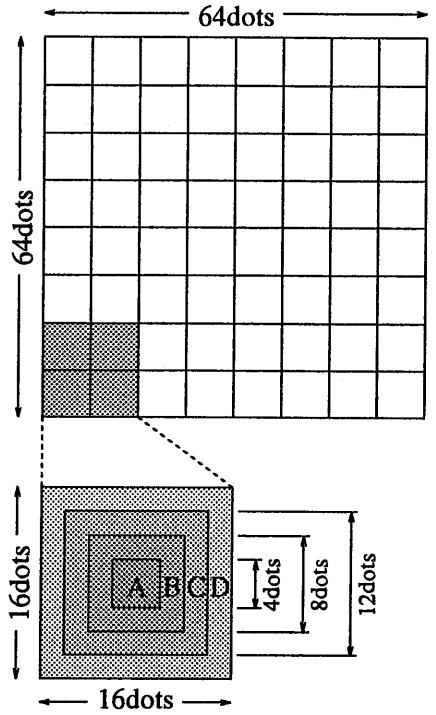
(a) イメージ　(b) ノイズ除去スムージング後　(b) 正規化後

(d) 細線化後　(e) 線素化後

図 2.3　各処理後のイメージ

2.3　特徴抽出

方向線素特徴量の抽出法を以下に記す．まず図 2.4左上のように，64 ドット×64 ドットの線素化された文字イメージの領域の縦と横とをそれぞれ 8 ドット間隔で分割する．次に，文字イメージの左上から 16 ドット×16 ドットを半分ずつ重複させ，図 2.5の様に左上から順に 1 から 49 の番号が付けられた小領域を作る．一つの小領域内に存在する各方向の線素の数を図 2.4右の様に重み付けして数えあげ，その小領域の特徴量とする．したがって，各小領域の特徴量は 4 次元のベクトルとなり，一つの文字の特徴量は 196(= 4 × 49) 次元のベクトルとなる．



方向線素：“｜”, “—”, “／”, “＼”
1 領域：16 × 16 ドット
領域数：7 × 7 = 49 個
次元数：4 × 49 = 196 次元
重みのかけ方：領域 A → 4
領域 B → 3
領域 C → 2
領域 D → 1

図 2.4　方向線素特徴量の抽出

1	3	5	7
15	17	19	21
29	31	33	35
43	45	47	49

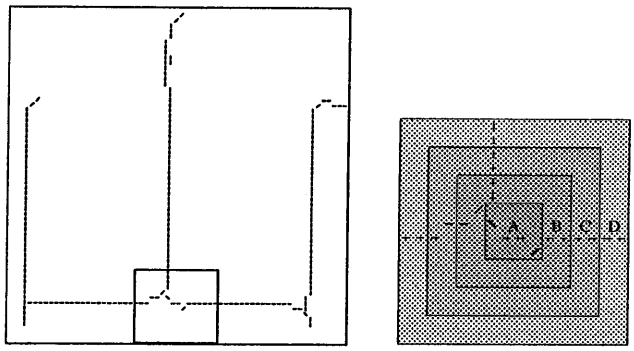
2	4	6
16	18	20
30	32	34
44	46	48

8	10	12	14
22	24	26	28
36	38	40	42

9	11	13
23	25	27
37	39	41

図 2.5　小領域の番号

図 2.6に「山」をサンプルにした 46 番目の小領域の特徴抽出例を示す．このサンプルでは，46 番目の小領域の特徴量は $F_{46} = (12, 29, 7, 4)$ となる．



$$F_{46\,|} = \overbrace{0 \times 4}^{A} + \overbrace{2 \times 3}^{B} + \overbrace{2 \times 2}^{C} + \overbrace{2 \times 1}^{D} = 12$$

$$F_{46-} = 2 \times 4 + 3 \times 3 + 4 \times 2 + 4 \times 1 = 29$$

$$F_{46\,/} = 1 \times 4 + 1 \times 3 + 0 \times 2 + 0 \times 1 = 7$$

$$F_{46\,\backslash} = 1 \times 4 + 0 \times 3 + 0 \times 2 + 0 \times 1 = 4$$

$$\mathbf{F_{46} = (12, 29, 7, 4)}$$

図 2.6　特徴抽出の例

これまで述べてきたことから方向線素特徴量の長所としては，

- 前処理の段階で正規化・細線化を行ったことにより，文字の大きさや線の太さの影響を受けにくい．
- 小領域を半分ずつ重複させ重み付けさせたことにより，位置ずれに強い．
- 部分情報を簡単に利用できる．

などが挙げられる．

第 3 章

部分領域における文字パターンの性質

3.1 まえがき

文字パターンは一定数の部品の組合せから成り，その一つ一つの部品を用いると，文字パターンは一つの符号で表すことができる．しかし，部首を部品とみなすとその切り出しは困難であることが知られている．そこで，ある一定の部分領域内に存在する文字パターンを各々の部品とみなすことにする．すると，部分領域の選び方が一つの問題となる．そこで本章では，部分領域を評価するための評価基準を導入し，どの様な部分領域がより有効に文字の特徴を反映するかを考察するために，部分領域における文字パターンの性質を調査する．

3.2 本研究における符号化の意味

文字パターンは一定数の部品の組合せから成ると考えることができる．文字パターンが P個の部品から成るとし，各部品が高々L 種類であるとすると，最大で L^P種類の文字パターンが存在する (実際の文字でないものも存在し得る)．この L と Pは，字種数を Kとするとき，$K \leq L^P$を満たしている．L 種類から成る部品 P個の組合せを考えることは，文字パターンが L 種類の記号を持つ長さ Pの記号列，つまり一つの符号で表わされることを意味する．本研究では，これを文字パターンの符号化と呼ぶ．

同一文字でもそのイメージは異なるのが普通であるが，異なるイメージであっても同一文字のイメージならば同一の符号に変換されることが望ましい．それを確認するため一つの文字に対して複数の文字イメージを用意する必要がある．本研究では全字種に対して各々一つだけのイメージの集合を文字セットと呼び，N個の文字セットを用意する．つまり，一つの文字に対して N個のイメージが存在することになる．

3.3 部分領域を選択するときの問題点

部分領域の選び方として，漢字の部首が存在する部分領域を積極的に選ぶことが一つの考え方である．例えば，偏を考慮に入れて文字イメージの左半分を部分領域とするとか，冠を考慮に入れて上側の 1/4 を部分領域とするなどが考えられる．ところが，部首を考慮に入れた場合いくつかの問題がある．例えば，偏が存在する部分領域は偏を持つ漢字に対しては有効であるが偏を持たない漢字に対しては有効でない，ということは充分に考えられる．あるいは，国構えが存在する部分領域を選ぶと，偏や冠が存在する部分領域がそれと重なってしまう．部分領域に重なりを持たせることが有効となるかも知れないが，重なった部分に存在するパターンは複数回用いられるので，処理が冗長となることも危惧される．また，偏の存在する部分領域を用いるにしても，どの程度の大きさが最適であるかを定めるための客観的基準はない．このように，部首を利用するにしても，どの部首を考慮すればよいかを判断する手立てがない．したがって，部首を考慮せずに部分領域を選ぶための客観的評価基準が必要となる．

いくつかの部分領域を選び，それぞれの部分領域において，複数の同一文字パターンが同じクラスに属するとする仮定する．この部分領域の組合せに基づいて文字の符号化を行えば，同一文字が同一符号に変換される．これは非常に望ましい符号化となるので，同一文字が同一符号に変換される様な状況を見付ける必要がある．そこで，部分領域において，複数の同一文字パターンが同じクラスに属する割合が高いほど値が高くなるような客観的評価基準である安定度を導入する．安定度で部分領域を評価することによって，部分領域を選択するときにどの様な領域を選択するべきかという問題を克服する．

3.4 安定度

どの様な部分領域が有効であるかを考察するために，安定度という評価基準を用いて部分領域を評価する．ここで安定度を「複数の同一文字パターンが同じクラスに属する割合」と定義する．以下では初めに，安定度を定式化するために，いままで文字パターンの一部の領域という意味で記してきた部分領域を厳密に定義し，次に，具体的に安定度を定式化する．

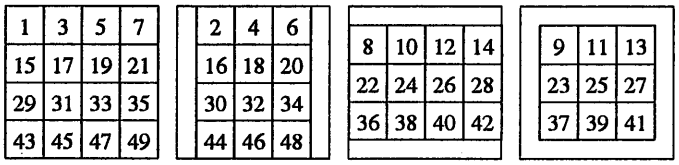
3.4.1 部分領域の定義

方向線素特徴量で表された文字パターンには，49 の小領域に対応したところにそれぞれ 4 次元のベクトルが存在し，各小領域におけるベクトルは文字パターンの局所的な情報を持っている．この小領域は文字イメージにおいては重なりをもっているが，特徴量としては各小領域ごとに分けて取り扱うことができる．そこで，方向線素特徴量の各小

領域を，図 3.1(b) の様に 49(= 7 × 7) 個のブロックで表し，図 3.1(a) で表される文字イメージの段階での小領域と対応させる．図 3.1(b) で表されるブロックの全ての集合を $\mathcal{U}$ で表し，部分領域を

$$\mathcal{P}_p \subseteq \mathcal{U},\ (p = 1, \ldots, P), \tag{3.1}$$

とする．ただし，Pは部分領域の数である．また，各部分領域間に重なりを許す．つまり，$\mathcal{P}_i \cap \mathcal{P}_j \neq \emptyset,\ (i \neq j)$ であってもよいものとする．また，部分領域の面積をそこに含まれる小領域の数と定義する．つまり，2 × 3 のブロックの集合で表される部分領域の面積は 6 である．



(a) 文字イメージの小領域の番号

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	32	33	34	35
36	37	38	39	40	41	42
43	44	45	46	47	48	49

(b) ブロックによる小領域の表現

図 3.1 小領域の番号とブロックの番号の対応

3.4.2 安定度の定式化

$k = 1, \ldots, K$，$n = 1, \ldots, N$に対して，k番目の字種の n セット目のサンプルの，部分領域$\mathcal{P}_p$における特徴ベクトルを部分ベクトルと呼び，$\boldsymbol{v}_{kn}(\mathcal{P}_p)$ と記す．この部分ベクトルに対する字種ごとの平均を標準部分ベクトルと呼び，k番目の字種の標準部分ベクトルを$\overline{\boldsymbol{v}}_k(\mathcal{P}_p)$ と記す．つまり，

$$\overline{\boldsymbol{v}}_k(\mathcal{P}_p) = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{v}_{kn}(\mathcal{P}_p), \tag{3.2}$$

と与えられる．$\{\overline{\boldsymbol{v}}_1(\mathcal{P}_p), \ldots, \overline{\boldsymbol{v}}_K(\mathcal{P}_p)\}$ を L 個のクラスタに分類する．各クラスタの平均値ベクトルを$\boldsymbol{c}_l(\mathcal{P}_p),\ (l = 1, \ldots, L)$ と記す．これらの平均値ベクトルを用いることで，k 番目の字種の標準部分ベクトル$\overline{\boldsymbol{v}}_k(\mathcal{P}_p)$ が属するクラス番号$\hat{l}_k(\mathcal{P}_p)$ は，

$$\hat{l}_k(\mathcal{P}_p) = \underset{l}{\operatorname{argmin}}\, d(\overline{\boldsymbol{v}}_k(\mathcal{P}_p), \boldsymbol{c}_l(\mathcal{P}_p)), \tag{3.3}$$

で与えられる．ここで，$d(\boldsymbol{u}, \boldsymbol{v})$ は二つのベクトル$\boldsymbol{u}$，$\boldsymbol{v}$間の距離を表し，$\underset{l}{\operatorname{argmin}}$はそれ以降の式が最小となる lを与える．部分ベクトル$\boldsymbol{v}_{kn}(\mathcal{P}_p)$ が属するクラス番号 $l_{kn}(\mathcal{P}_p)$ は

$$l_{kn}(\mathcal{P}_p) = \underset{l}{\operatorname{argmin}}\, d(\boldsymbol{v}_{kn}(\mathcal{P}_p), \boldsymbol{c}_l(\mathcal{P}_p)), \tag{3.4}$$

で与えられる．k番目の字種に対して標準部分ベクトルが属するクラスと同じクラスに属する部分ベクトルの数を $M_k(\mathcal{P}_p)$ で表す．つまり，

$$M_k(\mathcal{P}_p) = \sum_{n=1}^{N} \mathrm{delta}(\hat{l}_k(\mathcal{P}_p), l_{kn}(\mathcal{P}_p)), \tag{3.5}$$

と与えられる．ここで，

$$\mathrm{delta}(i,j) = \begin{cases} 1 & i = j. \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

である．全字種の $M_k(\mathcal{P}_p)$ を合計し，その全サンプル数に対する割合を安定度 $S(\mathcal{P}_p)$ と定義する．つまり，

$$S(\mathcal{P}_p) = \frac{1}{KN} \sum_{k=1}^{K} M_k(\mathcal{P}_p), \tag{3.7}$$

と与えられる．

式 (3.5)，式 (3.6)，式 (3.7) より明らかなように，$0 \leq S(\mathcal{P}_p) \leq 1$ であり，$S(\mathcal{P}_p)$ が 1 に近い部分領域では同一文字のパターンの変動が小さいことを意味する．

3.5 性質調査実験

部分領域における文字パターンの性質を調査するため，前節で定式化した安定度を求める実験を行う．

3.5.1 対象とする部分領域

対象とする部分領域は，図 3.1(b) のブロックの集合のうち，その形が長方形となるものとする．取り得る長方形の辺の長さは，縦が 1 から 7 の 7 通りであり，それぞれに対して横も 1 から 7 の 7 通りである．また，縦の長さが l_v，横の長さが l_hである部分領域は 7×7 の領域に $(8 - l_v)(8 - l_h)$ 個取ることができる．したがって，7×7 のブロックの集合内に取ることができる部分領域は全部で，

$$\sum_{l_v=1}^{7} \sum_{l_h=1}^{7} (8 - l_v)(8 - l_h) = 784(\text{個}), \tag{3.8}$$

となる．対象とする部分領域の例を図 3.2に示す．

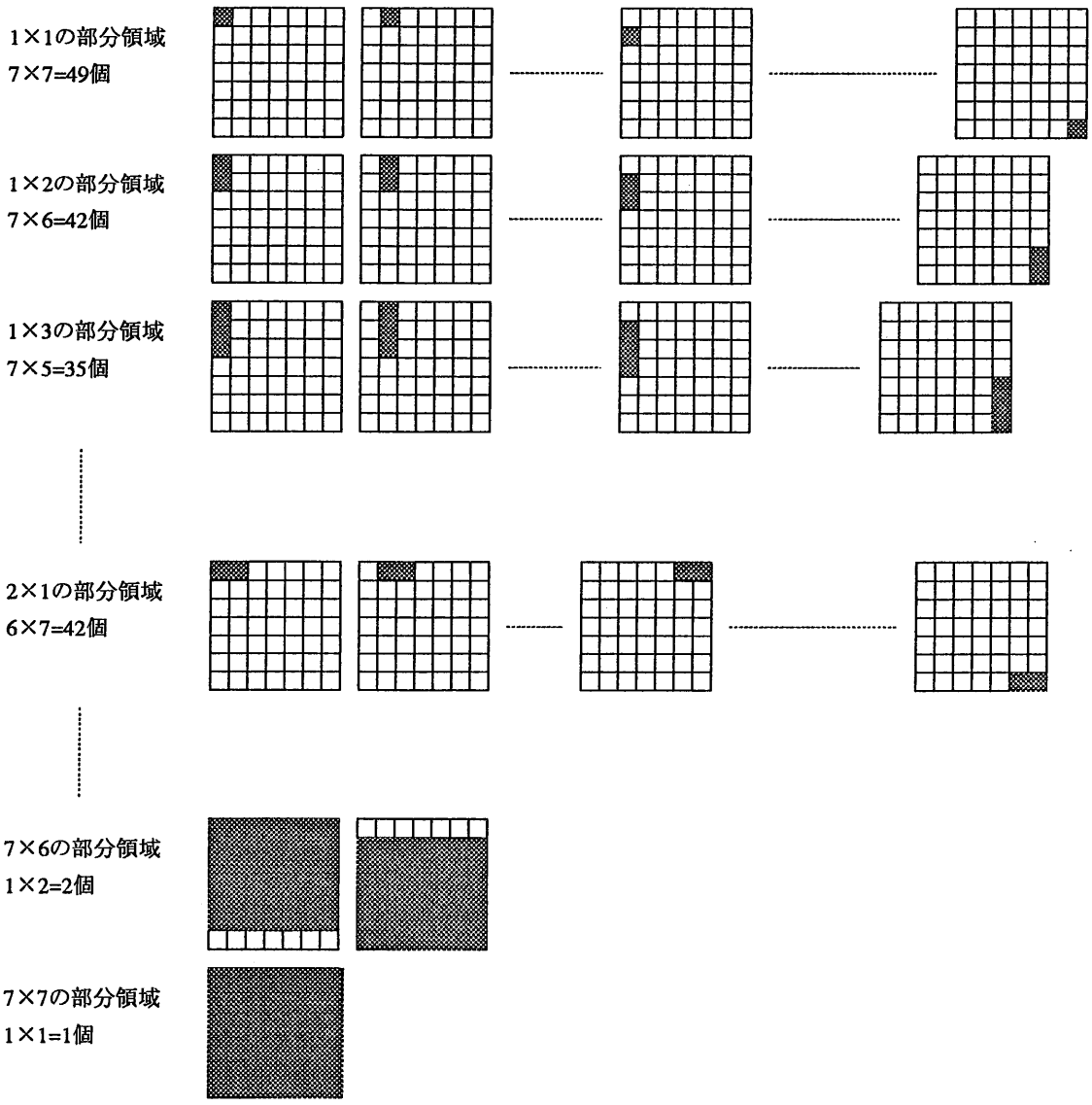


図 3.2　対象とする部分領域

3.5.2 実験条件

JIS 第 1 水準漢字 2965 字の印刷文字を対象とし，10 ポイントの明朝体フォントの文字セットを 5 セットを用意した．文字イメージは，400dpi のスキャナでとりこまれた二値画像から，文字ごとに正確に切り出されたものを使用した．クラスタリングの際のクラス数を 2，4，8，16，32，64 の 6 種類とし，それぞれのクラス数における安定度を求めた．クラス数を 2 のべき乗にしたのは，クラスタリング手法として LBG 法[8]を用いたためである．二つのベクトル間の距離として，ユークリッド距離の二乗を用いた．二つのベクトル$\boldsymbol{u} = (u_1, \ldots, u_m), \boldsymbol{v} = (v_1, \ldots, v_m)$ が与えられたとき，ユークリッド距離の二乗 $D_e(\boldsymbol{u}, \boldsymbol{v})$ は，

$$D_e(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{m} |u_i - v_i|^2, \tag{3.9}$$

で与えられる．ここで，m は各ベクトルの次元数である．

以上から本実験で用いた字種数 K，文字セット数 N，クラス数 L はそれぞれ次の様になる．

$$\begin{aligned} K &= 2965, \\ N &= 5, \\ L &= 2, 4, 8, 16, 32, 64 \text{ のいずれか一つ}. \end{aligned}$$

3.5.3 実験結果

図 3.3に各クラス数において安定度の高い順に部分領域を並べた様子を示す．全てのクラス数における安定度の平均値が高い順に並べたものを overall の欄に示す．また，表 3.1は図 3.3に示された部分領域の安定度を表している．

表 3.1から，安定度が 1 になる部分領域，つまり，複数の同一文字パターンが必ず同一クラスに属するような部分領域は存在しないことが分った．したがって，同一文字パターンであってもパターンの変動があり，対象とした部分領域では同一文字パターンの変動を完全に吸収できないことが分った．あるいは，どのような部分領域を選んでも安定度は 1 にならないということも充分に考えられる．

クラス数との関係を考えると，クラス数が増加するにつれ面積の大きい部分領域が安定度の高い順位に現れることが分る．この理由は以下の様に考えられる．面積の小さい部分領域においては部分ベクトルの次元数も小さいので，クラス数が大きくなるとベクトルの成分ごとの小さな差の影響が増し，同一クラスに属することができるサンプルが減少する．しかし，面積の大きい部分領域においては部分ベクトルの次元数も大きく，ベクトルの成分ごとの差の影響が増えてもベクトル全体ではその影響が緩和されるため，

次元数の小さいサンプルに較べると，同一クラスに属するサンプル数の減少が抑えられるためである．

また，全体的に見て，文字の左側の部分領域の安定度が高いことが分る．これは漢字の偏が存在する部分領域であり，方向線素特徴量で表された漢字に関しては，偏の存在する部分領域において複数の同一文字パターンが同一クラスに属する割合が高いことが分った．偏という知識を与えていないにも拘らず，統計的にこういった結果が得られたことは興味深いことである．これは，漢字の場合，部首領域を積極的に利用することの有効性を示すものと言える．

この様な結果が得られたので，漢字の部首を考慮した部分領域を意図的に作り，同様に安定度を調査したところ，表 3.2の様な結果が得られた．表 3.2の部分領域は左からそれぞれ，門構え，門構えと冠の中間，冠，脚とにょうの中間，にょう，垂れ (やまいだれなどを考慮して左部分の幅が大きい)，垂れ (まだれなどを考慮して左部分の幅が小さい) を意図したものである．また，安定度の右に括弧で括られた数字があるものは，括弧内の数字が図 3.3における順位を表している．但し，この数字はその部分領域が単独で順位付けされたときのものである．そのため，表 3.2の同じクラス数の欄において，安定度が異なるにも拘らず括弧内に同じ数字が入ることもある．表 3.2を見ると，クラス数が 16 のとき門構えの部分領域の安定度が最高になっていることが分る．このためこの部分領域は文字パターンの特徴を良く表していると言える．順位が 10 位以内に入った他の部分領域 (にょう，垂れ) については，図 3.3において比較的安定度の高い左側の部分領域に横向きの部分領域が足されているものであり，左側の部分領域の安定度の高さによる影響が大きいだけかも知れない．左側の部分領域の安定度の高さによる影響が大きいことは，冠を考慮した部分領域の安定度があまり高くないことからも窺える．このため，本当にこれらの部首が有効であったのか，それとも，単に左側の部分領域の安定度の影響が大きいのかを調べるために，さらなる調査が必要であると思われる．

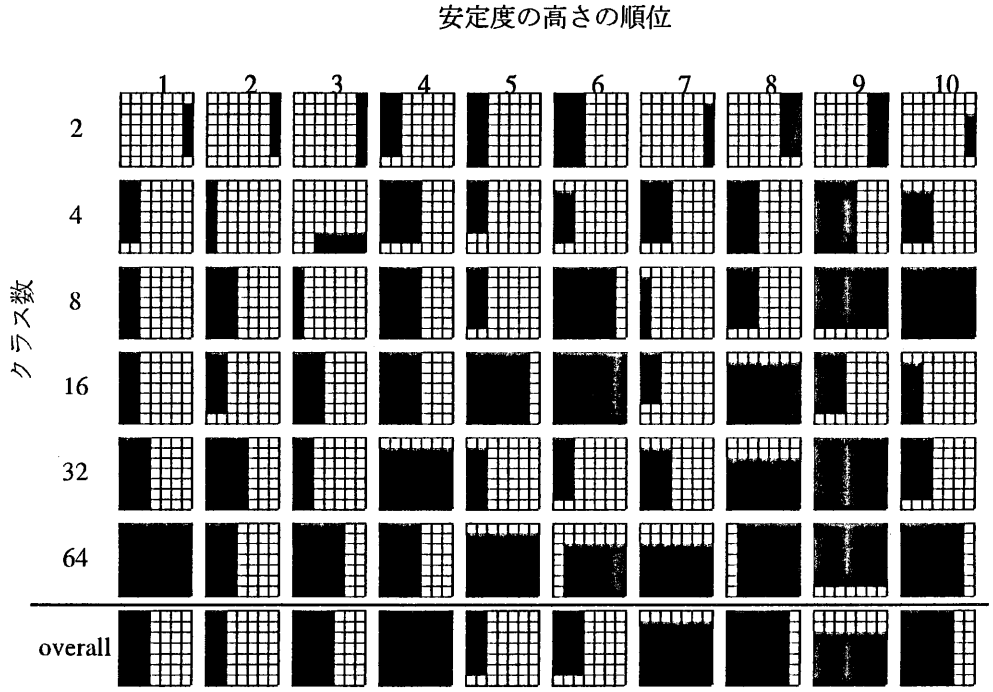


図 3.3　安定度の高い部分領域

表 3.1　安定度の高い部分領域の安定度

クラス数	安定度の高さの順位									
	1	2	3	4	5	6	7	8	9	10
2	0.9773	0.9756	0.9750	0.9746	0.9734	0.9730	0.9728	0.9725	0.9720	0.9717
4	0.9398	0.9392	0.9359	0.9357	0.9353	0.9352	0.9343	0.9342	0.9338	0.9332
8	0.9222	0.9217	0.9184	0.9155	0.9144	0.9139	0.9130	0.9120	0.9118	0.9110
16	0.9060	0.9031	0.9026	0.9000	0.8996	0.8972	0.8958	0.8942	0.8935	0.8920
32	0.9046	0.9010	0.8987	0.8964	0.8950	0.8940	0.8932	0.8927	0.8924	0.8924
64	0.9154	0.9062	0.9055	0.9032	0.8996	0.8990	0.8985	0.8984	0.8980	0.8976
overall	0.9237	0.9212	0.9206	0.9187	0.9163	0.9156	0.9142	0.9133	0.9133	0.9123

表 3.2　部首を考慮した部分領域の安定度

クラス数	部分領域						
							
2	0.9737(5)	0.9685	0.9520	0.9701	0.9687	0.9724(9)	0.9753(3)
4	0.9297	0.9276	0.9195	0.9302	0.9242	0.9256	0.9245
8	0.9150(5)	0.9048	0.8956	0.9022	0.9177(4)	0.9175(4)	0.9021
16	0.9061(1)	0.8902	0.8855	0.8936(9)	0.8957(8)	0.8902	0.8988(6)
32	0.8879	0.8831	0.8730	0.8799	0.8874	0.8937(7)	0.8913
64	0.9026(5)	0.8884	0.8706	0.8845	0.8923	0.9010(5)	0.8969
overall	0.9192(4)	0.9104	0.8994	0.9101	0.9143(7)	0.9167(5)	0.9148(7)

3.6　まとめ

本章では，文字パターンを部品の組合せと考え，様々な部分領域から有用な部分領域を選定するための評価基準として安定度を導入した．実験により以下のことを確かめた．安定度が 1 となる部分領域は存在しない．クラス数が増加するにつれ，面積の大きい部分領域が安定度の高い順位に頻繁に現れる．全体的に見て文字の左側の部分領域の安定度が高い．つまり，方向線素特徴量で表された漢字のパターンは左側の部分が比較的安定である．門構えを意図した部分領域の安定度が比較的高い．垂れやにょうを意図した部分領域の安定度も比較的高いが，左側の部分領域の安定度の高さの影響によるものかも知れない．以上のうち最後の結果についてはさらなる調査が必要である．

第4章

文字パターンの符号化

4.1 まえがき

本研究で考察する符号化は，複数の部分領域を組合せ，それを基に文字パターンを符号化する．そこでまず最初に，どの様に部分領域を組合せるとより有効な符号化を行うことが可能であるかについて考察する．次に，この部分領域の組合せを基にした符号化法について考察する．最後に，符号の一致性に基づく大分類の可能性を検討し，本研究で提案する文字パターンの符号化法の性能を評価する．

4.2 部分領域の組合せ

4.2.1 部分領域を組合せるときの問題点

同一文字の符号の一致性を高めるためには，できるだけ安定度が高く，また，符号として有効となる部分領域を選ぶことである．有効な符号化を与える部分領域は単に安定度が高いだけでなく，ある決った面積を持つものと考えられる．また，各部分領域どうしに重なりを許すかどうかといった問題がある．そこで，この二つの問題について考察し，部分領域を組合せる方法について説明する．あらかじめ部分領域は安定度の高い順にソートしてあるものとする．

初めに部分領域の面積について考える．1×1 の部分領域など面積が小さい部分領域では安定度が低く符号化には有効でないので，符号化に用いる部分領域の面積に下限を設ける．これにより，同一文字の符号の一致性を低めるという状況を避ける．また，あまり面積の大きい部分領域を用いると文字パターンを覆う部分領域の組合せが少なくなり，作成される符号は少数となる．このため，一つの符号が多くの文字に対応することになる．さらに，ある程度以上に面積の大きい部分領域では，文字パターンの局所的な情報を得ることができないと考えられる．したがって，部分領域の面積に上限も設けることとする．

次に考えることは，部分領域どうしの重なりを許すか否かである．安定度の高い部分領域は比較的面積が大きので，面積の大きい部分領域を用いることは利にかなっているが，そうすると文字パターンの残りの領域が小さくなる．部分領域の重なりを許さないと，残りの領域の部分領域の組合せが限定され，安定度の低い部分領域を用いなければならない状況になる可能性もある．また前章で，漢字の偏の存在する部分領域は非常に安定度が高いことが分っているので，偏の存在する部分領域 (つまり文字パターンの左側の部分) を用いると，偏ではなく脚を持つ漢字のために脚の存在する部分領域 (つまり文字パターンの下側の部分) を用いようとしても，偏のある部分領域と重なってしまい選ぶことができなくなる．よって，多くの文字パターンに対して柔軟性を持たせるために，部分領域の重なりを許すこととする．

実験によりどの様な部分領域の組合せが符号化に有効であるかについて以降でより詳細に考察する．

4.2.2 部分領域を組合せる手順

まず実験に用いる部分領域の組合せについてであるが，前に述べたように，

- 部分領域の安定度
- 部分領域の面積
- 部分領域どうしの重なり

を考慮に入れ，以下の流れで部分領域の組合せを選ぶ．

STEP1 安定度の高い順に部分領域 $\mathcal{P}_r$, $(r = 1, \ldots, 784)$ をソートし，ソートした結果の順位を部分領域の新たなインデックスとする．

$$\mathcal{P}_1, \ldots, \mathcal{P}_r, \ldots, \mathcal{P}_{784}.$$

STEP2 $r \leftarrow 0$. $\mathcal{M} \leftarrow \emptyset$.

STEP3 $r \leftarrow r + 1$.

STEP4 以下の 2 式が満たされれば **STEP5** へ，満たされなければ **STEP3** へ．

$$A_{\min} \leq \mathrm{area}(\mathcal{P}_r) \leq A_{\max}. \tag{4.1}$$

$$\mathrm{area}(\mathcal{P}_r \cap \mathcal{M}) \leq R \cdot \mathrm{area}(\mathcal{P}_r). \tag{4.2}$$

STEP5 $\mathcal{M} \leftarrow \mathcal{M} \cup \mathcal{P}_r$.

STEP6 area($\mathcal{M}$) < 49 かつ $r < 784$ ならば **STEP3** へ.

STEP7 終了.

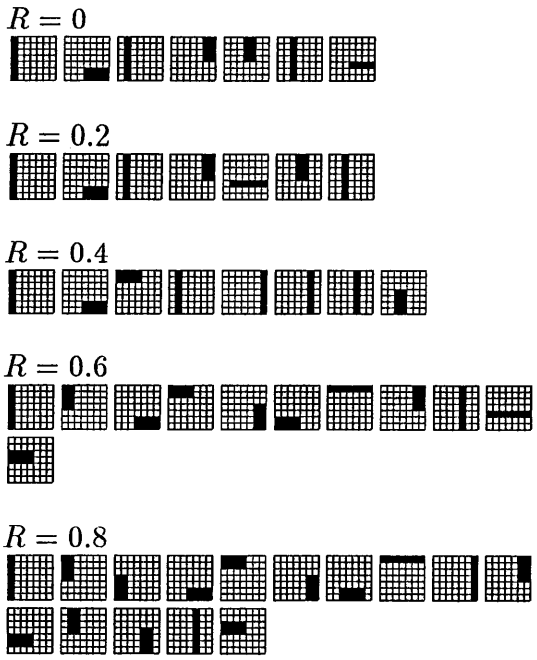
各記号はそれぞれ以下のものを表す.

$\mathcal{M}$: 既に部分領域によって覆われた領域.
area($\mathcal{A}$) : 領域 $\mathcal{A}$ の面積.
$A_{\min}$: 部分領域の面積の上限.
$A_{\max}$: 部分領域の面積の下限.
R : $\mathcal{M}$ と部分領域の重なりの割合を規定する定数.

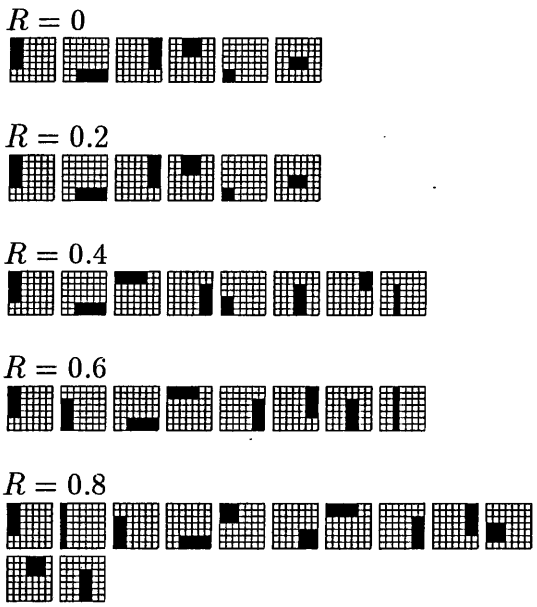
重なりを規定する定数 Rについては，$0 \leq R \leq 1$ であり，$R = 0$ のときは部分領域の重なりを全く許さないことを意味し，$R = 1$ のときは部分領域はいくらでも重ねられることを意味する.

4.2.3 部分領域を組合せる実験

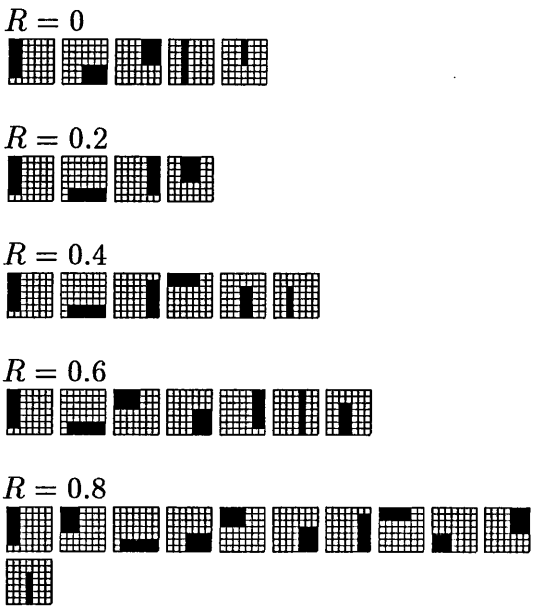
部分領域の安定度として，前章で求めた安定度のうち，全てのクラス数における安定度の平均値，つまり，図 3.3の overlap の欄の安定度を用いた．これは，クラス数を決められない段階での実験のため，総合的な安定度を用いるべきであると考えたためである．$A_{\min}$を 4 に固定し，$A_{\max}$を 8，10，12，14 に変化させた．各 $A_{\max}$に対して Rを 0，0.2，0.4，0.6，0.8 の値に変化させた．したがって，この場合得られる部分領域の組合せは 20 種類となる．この条件の下で得られた部分領域の組合せを図 4.1に示す．図 4.1を見ると，許される部分領域の重なりが大きくなるほど組合される部分領域の数が増えることが分かる．これは，許される重なりが大きくなると，一つの部分領域によって新らしく覆われる領域が小さくてもよくなり，全領域を覆うのに多くの部分領域が必要となるためである．また，面積の上限が大きくなるほど組合される部分領域の数が減ることが分かる．これは，面積が大きくなると，一つの部分領域で覆われる領域が大きくなり，少数の部分領域で全領域を覆うことができるためである．したがって，組合される部分領域の数を多くしたければ，できるだけ部分領域の面積を小さく，部分領域どうしの重なりを大きくすればよい．逆に組合される部分領域の数を少なくしたければ，できるだけ部分領域の面積を大きく，部分領域どうしの重なりを小さくすればよい．また，図 4.1(b) の $R = 0$ と $R = 0.2$ では，全く同じ組合せとなっている．これは，部分領域の面積の上限も許される重なりも小さいため，他の部分領域が選ばれなかったためであ



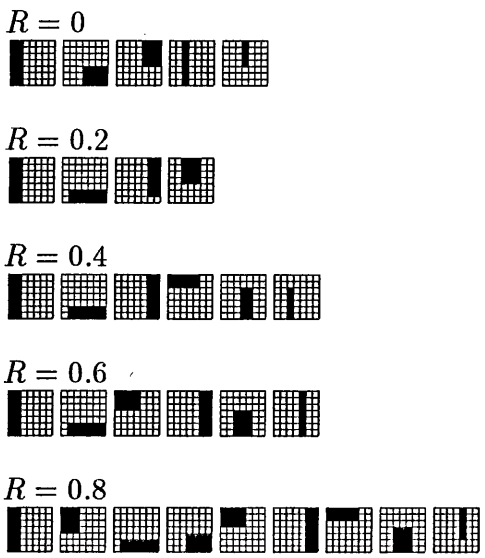
(a) 面積が 4 以上 8 以下



(b) 面積が 4 以上 10 以下



(c) 面積が 4 以上 12 以下



(d) 面積が 4 以上 14 以下

図 4.1　得られた部分領域の組合せ

ると考えられる．このように，組合せる条件の差が小さいと部分領域の組合せが全く同じになることも有り得る．

4.3 符号化法

部分領域の組合せを基にした符号化を行う．この符号化は後に述べる大分類のために行うので，未知入力文字パターンの符号化が主となる．そのため，

- 未知入力文字パターンの符号化に必要な辞書を作成し，
- それを基に未知入力文字パターンを符号化し，
- 得られた符号の評価をする，

ということが必要となる．そして，符号化に要する計算量は大分類の性能を定める一つの要素であるので，計算量についても言及する．

4.3.1 辞書作成

未知入力文字パターンが符号に変換されても，単に符号を得るだけではその符号の評価ができない．そのため，あらかじめ，標準符号を作成する必要があり，それと比較することで符号を評価する．また，未知入力文字パターンを符号化するためには，標準符号の作成に用いた部分領域の情報と，その部分領域におけるクラス分けのための情報が必要となる．したがって，

- 標準符号
- 部分領域の情報
- クラス分けのための情報

を保存するものが必要となる．本研究ではこれを辞書と呼ぶ．以下に辞書作成の手順を示す．

STEP1 P個の部分領域の組合せ $\{\mathcal{P}_1,\ldots,\mathcal{P}_P\}$ を選ぶ．この P個の部分領域の情報を辞書に保存する．部分領域の情報とは，各部分領域がどのような小領域の集合から成るかを記述したものである．

STEP2 $p=1,\ldots,P$に対して以下のことを行なう．言葉の定義やベクトル間の距離およびクラスタリング手法は，安定度を求めたときと同様である．

1. $\mathcal{P}_p$における全サンプルの部分ベクトル$\boldsymbol{v}_{kn}(\mathcal{P}_p)$ を求める.
2. 各文字の標準部分ベクトル$\overline{\boldsymbol{v}}_k(\mathcal{P}_p)$ を求める.
3. $\{\overline{\boldsymbol{v}}_1(\mathcal{P}_p),\ldots,\overline{\boldsymbol{v}}_K(\mathcal{P}_p)\}$ を L 個のクラスタに分類する.
4. 各クラスタの平均値ベクトルを$\boldsymbol{c}_l(\mathcal{P}_p)$, $(l=1,\ldots,L)$ と記す. これらを $\mathcal{P}_p$における代表ベクトルと呼び, L 個の代表ベクトルを辞書に保存する. つまり先に述べたクラス分けのための情報とは代表ベクトルのことである.

STEP3 $p=1,\ldots,P$, $k=1,\ldots,K$に対して, $\overline{\boldsymbol{v}}_k(\mathcal{P}_p)$ の属するクラス番号$\hat{l}_k(\mathcal{P}_p)$ を次式で決定し, KP個のクラス番号の集合 $\{\hat{l}_k(\mathcal{P}_p)\}$ を辞書に保存する.

$$\hat{l}_k(\mathcal{P}_p)=\underset{l}{\operatorname{argmin}}\, d(\overline{\boldsymbol{v}}_k(\mathcal{P}_p),\boldsymbol{c}_l(\mathcal{P}_p)). \tag{4.3}$$

STEP4 終了.

辞書内では k番目の字種は P個の記号の列

$$(\hat{l}_k(\mathcal{P}_1),\hat{l}_k(\mathcal{P}_2),\ldots,\hat{l}_k(\mathcal{P}_p)), \tag{4.4}$$

と対応している. これを k番目の字種の標準符号とする.

4.3.2 未知入力文字パターンの符号化

作成された辞書によって未知入力文字パターンを符号化する方法を述べる. 未知入力文字パターンの符号化の流れを図 4.2に示す. 辞書内の各部分領域の情報から未知入力文字パターンの部分ベクトル$\boldsymbol{v}_x(\mathcal{P}_p)$ を求め, 辞書内の代表ベクトルとの整合を行い, 次式で $l_x(\mathcal{P}_p)$ を決定する.

$$l_x(\mathcal{P}_p)=\underset{l}{\operatorname{argmin}}\, d(\boldsymbol{v}_x(\mathcal{P}_p),\boldsymbol{c}_l(\mathcal{P}_p)). \tag{4.5}$$

$p=1,\ldots,P$に対して整合を行い, 符号

$$(l_x(\mathcal{P}_1),l_x(\mathcal{P}_2),\ldots,l_x(\mathcal{P}_P)), \tag{4.6}$$

を得る.

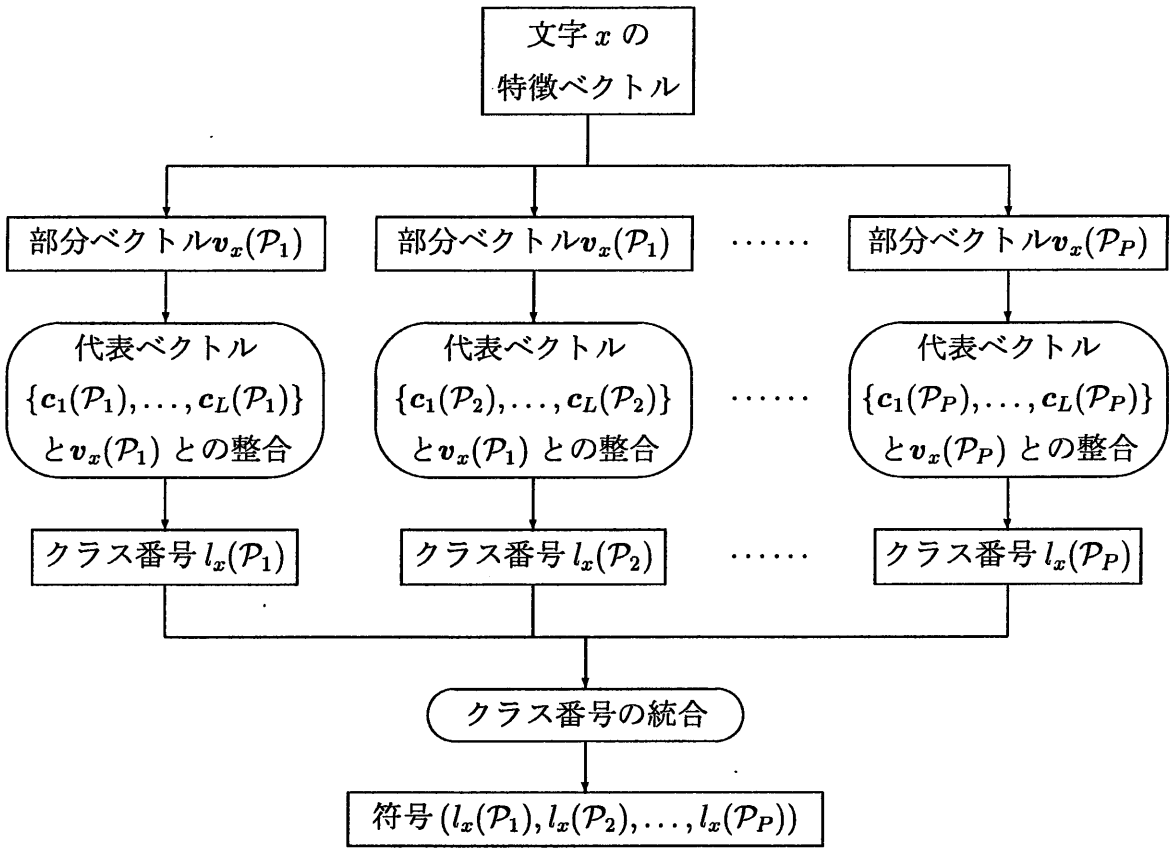


図 4.2　未知入力文字パターンの符号化法

4.3.3　符号の分類

未知入力文字パターンの符号を標準符号と比較したとき，次の 2 通りを考える．

(I) 正解文字に対応する標準符号に一致する．

(II) 正解文字でない文字に対応する標準符号に一致する．

ここで，正解文字とは未知入力文字パターンの実際の文字である．(I)，(II) に対して符号は以下の 4 種類の符号に分類される．

不一致符号　：　(I) と (II) のいずれも満たさないもの．
誤り符号　：　(I) を満たさず (II) を満たすもの．
リジェクト符号　：　(I) と (II) の両方とも満たすもの．
正解符号　：　(I) を満たし (II) を満たさないもの．

4.3.4 符号化に必要な計算量

未知入力文字パターンは式 (4.5) と式 (4.6) に基づき符号に変換される．式 (4.5) の右辺の距離計算の計算量は，部分領域における部分ベクトルの次元数に比例する．その距離計算をクラス数だけ行わねばならないので，一つの部分領域においてパターンが属するクラス番号を決定するのに必要な計算量は，部分ベクトルの次元数 $D(\mathcal{P}_p)$ とクラス数 L の積に比例する．対象となる全ての部分領域において式 (4.5) の左辺を求める必要があるので，符号を得るのに必要な計算量 T_{coding} は，

$$T_{\mathrm{coding}} = \sum_{p=1}^{P} a \cdot L \cdot D(\mathcal{P}_p), \tag{4.7}$$

である．a は比例定数である．

4.4 符号化による大分類

本研究では大分類を以下の様に行う．未知入力文字パターンを符号化し，得られた符号と一致する符号を辞書内から探す．一致した符号に対応する文字の集合をこの未知入力文字パターンの候補文字集合とする．

この符号による大分類を実現するための条件は，候補文字集合に必ず正解文字が含まれることである．つまり，未知入力文字パターンの符号が，正解文字に対応する辞書内の符号と一致しなければならない．また，満足しなければならない条件ではないが，大分類の後の処理量を低減するために，候補文字集合の要素数ができるだけ少ないことが望まれる．

そこで，符号による大分類の可能性を確かめるため，文字パターンの符号化を行い，符号の一致性を調べる．まず，一つの文字に対して一つだけの符号が対応する単一符号による実験，次に，一つの文字に対して複数の符号が対応する複数符号による実験の順に述べる．

4.4.1 単一符号による大分類実験

大分類の条件を満たすのは，**4.3.3**で分類された符号のうちリジェクト符号と正解符号の二つである．つまり，得られた符号がリジェクト符号ならば，それが正解文字に対応する符号であることは確かであり，正解符号ならば，それに加えて不正解文字に対応する符号のいずれとも一致しない．いずれにせよ，正解文字に対応する符号には一致するので，大分類が実現できることになる．

実験の目的

4.2.3で求めた部分領域の組合せ方で符号化するときに，どの組合せが有効かを考察する必要がある．符号化による大分類を実現するための条件は正解文字に対応する符号と一致することである．また，大分類の処理の低減化のために，符号を得るのに必要な計算量は小さい方が良い．そこで，計算量と符号の一致性の関係を調べ，符号化に適する部分領域の組合せを検討することを目的として実験を行う．

実験条件

辞書作成には安定度を求めたときと同じ 5 つの文字セットを用いる．辞書作成に用いない文字セットを用意し，その文字パターンを未知入力文字パターンとする．このセットは，10 ポイントの明朝体フォントの文字セットを 400dpi のスキャナで取り込み，二値化されたイメージから文字ごとに正確に切り出されたものである．部分領域としては**4.2.3**で得られた 20 種類の組合せを用いる．クラス数は組合された全ての部分領域で一定とし，その値を 2，4，8，16，32，64 に変化させる．

実験結果

図 4.3に，大分類を実現できた符号の割合を示す．以後この割合を分類率 (classification rate) と呼ぶ．横軸は部分領域の組合せから計算される計算量の最大値を 1 として正規化した値である．このグラフを見ると，分類率が最も高かった部分領域の組合せは，$A_{\max} = 12, R = 0.2$ のときのものであり，$L = 2$ のときに約 82%の分類率が得られている．このとき組合されている部分領域の数は 4 つであるので，標準符号の数は $2^4 = 16$ だけである．つまり，この部分領域の組合せによって全字種を 16 のクラスに大分類したとき，全字種の約 82%が正しく分類できたことになる．

図 4.4に，全ての字種数に対する正解符号数の割合を示す．以後この割合を正解率 (correct rate) と呼ぶ．正解率が最も高かった部分領域の組合せは，$A_{\max} = 14, R = 0.2$ のときのものである．この部分領域の組合せのときの正解率は実線で表されており，$L = 64$ のときに最高の正解率が得られている．この部分領域の組合せは部分領域の面積の上限が最大 ($A_{\max} = 14$) のときに得られたもので，かつ，組合される部分領域の個数は得られた全ての組合せの中で最小である．横軸の値が同程度である部分領域の組合せとして，$A_{\max} = 8, R = 0.0$ のときのものを点線で表す．この部分領域の組合せでは $L = 8$ のときに正解率が最高になる．このときの計算量では $A_{\max} = 14, R = 0.2$ のときのものに較べ正解率は高い．ところが，$L = 64$ のときには $A_{\max} = 14, R = 0.2$ のときのものに較べると，正解率は約半分である．したがって，同じ計算量という条件で評価するとき，ある程度計算量を大きくできるのならば，部分領域の面積を大きくとり，部分領

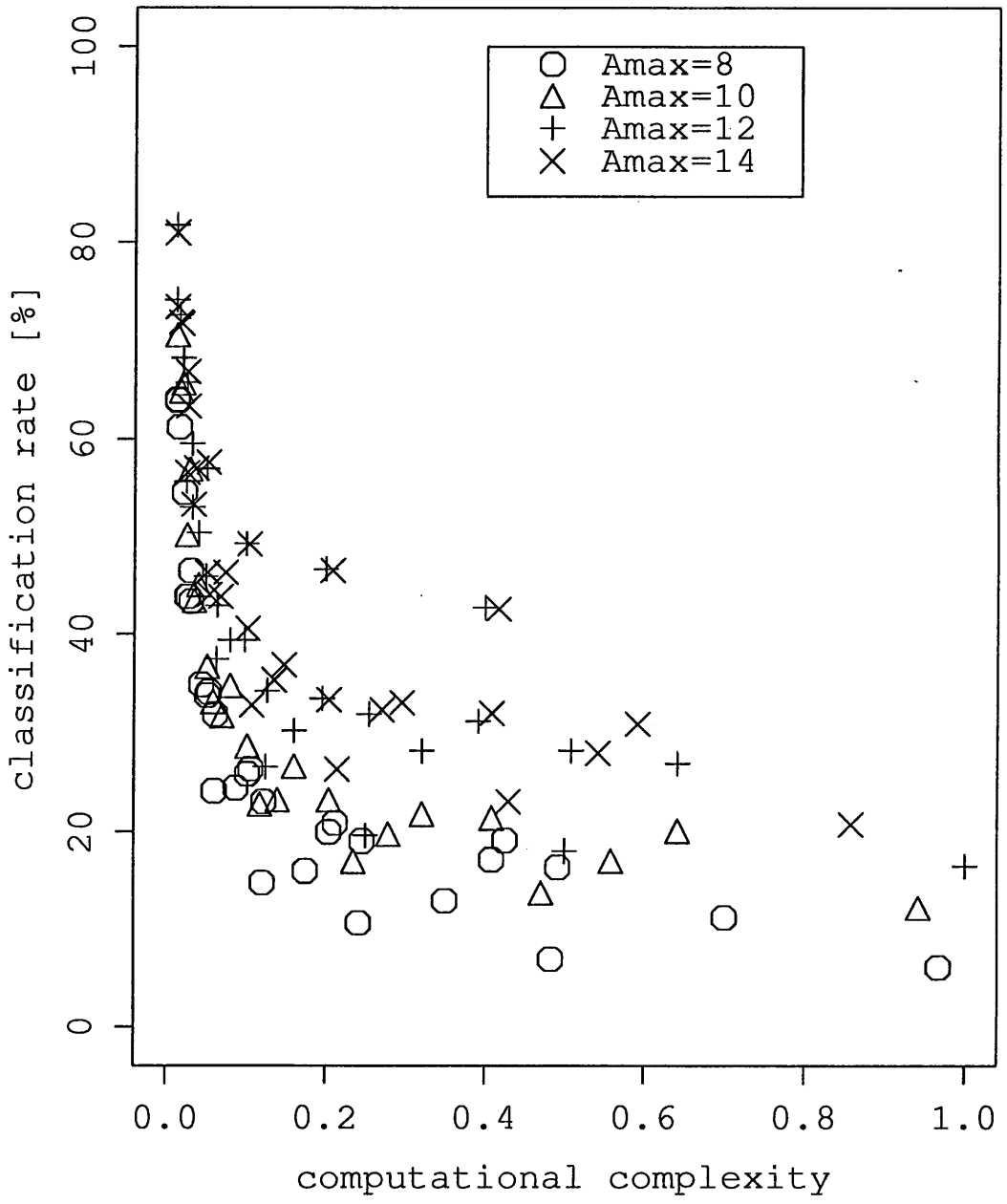

Amax=8
Amax=10
Amax=12
Amax=14
classification rate [%]
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
computational complexity

図 4.3　計算量と分類率

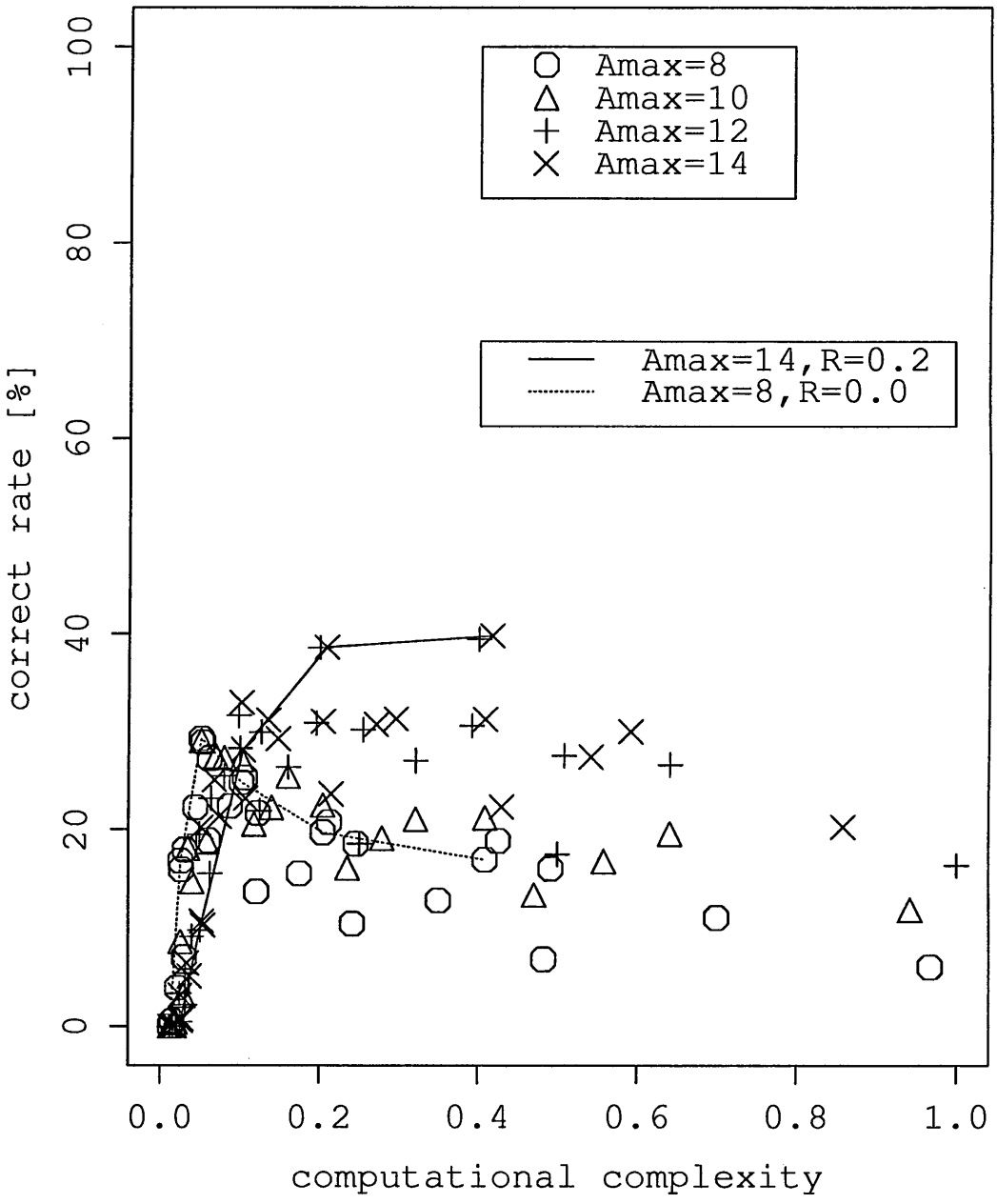

Amax=8
Amax=10
Amax=12
Amax=14
Amax=14,R=0.2
Amax=8,R=0.0
correct rate [%]
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
computational complexity

図 4.4　計算量と正解率

域の数を小さくする組合せを選ぶほうが高い正解率が得られると考えられる．

二つのグラフから，大分類の実現のためには，計算量が少ない部分領域の組合せの方が良いが，望ましい大分類を実現するにはある程度の次元数が必要であることが分る．

しかし，この程度での分類率では大分類法としては役立たないので，より高い符号の一致性を期待できる複数符号を次に述べる．

4.4.2　複数符号による大分類実験

各文字に対して一つの標準符号を用意する符号化では大分類の実現が困難であることが分ったので，より一致性の高い複数符号を導入する．これは，各部分領域において同一文字パターンの属するクラスを複数とし，未知入力文字パターンの属するクラスがその複数のクラスのいずれかと等しければ，その部分領域でのクラス分けが成功したとして，符号の一致性を高めようという考えの下に考案した．

実験の目的

複数符号を作成すること，および，未知入力文字パターンの符号との一致性を調べ，その有効性を確かめることを目的として実験を行う．

実験方法

複数の標準符号を作るため，**4.3.1**で示した辞書作成の方法の **STEP3** を以下の様に変更する．

STEP3′ $p=1,\ldots,P$，$k=1,\ldots,K$，$n=1,\ldots,N$に対して，$\boldsymbol{v}_{kn}(\mathcal{P}_p)$ の属するクラス番号 $l_{kn}(\mathcal{P}_p)$ を次式で決定し，KNP個のクラス番号の集合 $\{l_{kn}(\mathcal{P}_p)\}$ を辞書に保存する．

$$l_{kn}(\mathcal{P}_p)=\operatorname*{argmin}_{l} d(\boldsymbol{v}_1(\mathcal{P}_p),\boldsymbol{c}_l(\mathcal{P}_p)). \tag{4.8}$$

このとき，辞書内では k番目の字種は N行 P列の行列

$$\begin{pmatrix} l_{k1}(\mathcal{P}_1) & l_{k1}(\mathcal{P}_2) & \cdots & l_{k1}(\mathcal{P}_P) \\ \vdots & \vdots & \vdots & \vdots \\ l_{kN}(\mathcal{P}_1) & l_{kN}(\mathcal{P}_2) & \cdots & l_{kN}(\mathcal{P}_P) \end{pmatrix} \tag{4.9}$$

と対応している．これを k番目の字種の標準符号とする．

未知入力文字パターン x の符号を $(l_x(\mathcal{P}_1),\ldots,l_x(\mathcal{P}_P))$ とする．次式が成立したとき，未知入力文字パターン x の符号と k番目の字種の標準符号が一致したこととする．

$$l_x(\mathcal{P}_p) = l_{kn}(\mathcal{P}_p),\ 1 \leq \forall p \leq P,\ 1 \leq \exists n \leq N. \tag{4.10}$$

辞書作成に用いた文字セットは単一符号のときと同じ 5 セットである．したがって，標準符号は 5 行 P 列の行列となる．その他の実験条件は単一符号のときと同じである．

実験結果

図 4.5に，計算量と分類率の関係を示す．複数符号により単一符号よりも分類率は上っていることが分る．しかし，これでも 100%の分類率が得られていない．分類率が最も高かった部分領域の組合せは，$A_{\max} = 12, R = 0.2$ のときのものであり，$L = 2$ のときに約 96%の分類率が得られている．これは単一符号のときに最高の分類率を得たのと同じ組合せである．

図 4.6に，計算量と正解率の関係を示す．やはり単一符号のときと較べて正解率は向上している．正解率の最も高かった部分領域の組合せは，$A_{\max} = 14, R = 0.2$ のときのものであり，これは単一符号のときと同じ組合せである．このときの正解率の変化は実線で表されており，$L = 64$ のときに約 65%の正解率が得られている．単一符号のときと同様に，$A_{\max} = 8, R = 0.0$ のときの正解率の変化を点線で表す．このとき $L = 16$ で最高の正解率が得られている．また，単一符号のときと同様に計算量の増加につれて正解率の逆転が起こっている．したがって単一符号と同様に，ある程度計算量を大きくできるならば，部分領域の面積を大きくとり部分領域の数を小さくする組合せを選ぶほうが高い正解率が得られる．

4.4.3 大分類実験の考察

2 種類の符号による実験結果から言えることは以下の様なことである．大分類を実現するには計算量が小さくなる様な部分領域の組合せを選べばよい．ただし，分類数が小さいので望ましい大分類とは言えない．計算量が小さいときは小さな部分領域の組合せを用いたほうが正解率は高いが，計算量がある程度大きくなると大きな部分領域の組合せを用いたほうが正解率は高くなる．

しかし，この方法では完全な大分類には至っていない．この理由は以下の様に考えられる．まず，安定度が 1 となる部分領域が存在しないことである．このため，一つの部分領域において同一文字パターンの属するクラスを複数にしても，未知入力文字パターンの属するクラスがそれらのいずれとも異なることがあり，それだけで符号は一致しない．また，複数の部分領域を組合せているので，一つの部分領域だけで属するクラスが異なっても標準符号と未知入力文字パターンの符号は一致しない．したがって，部分領域の数は少ないほうがよいと考えられる．結局，一つの部分領域で大分類することにより，細かく分類することはできないが，最高の分類率が得られるものと思われる．した

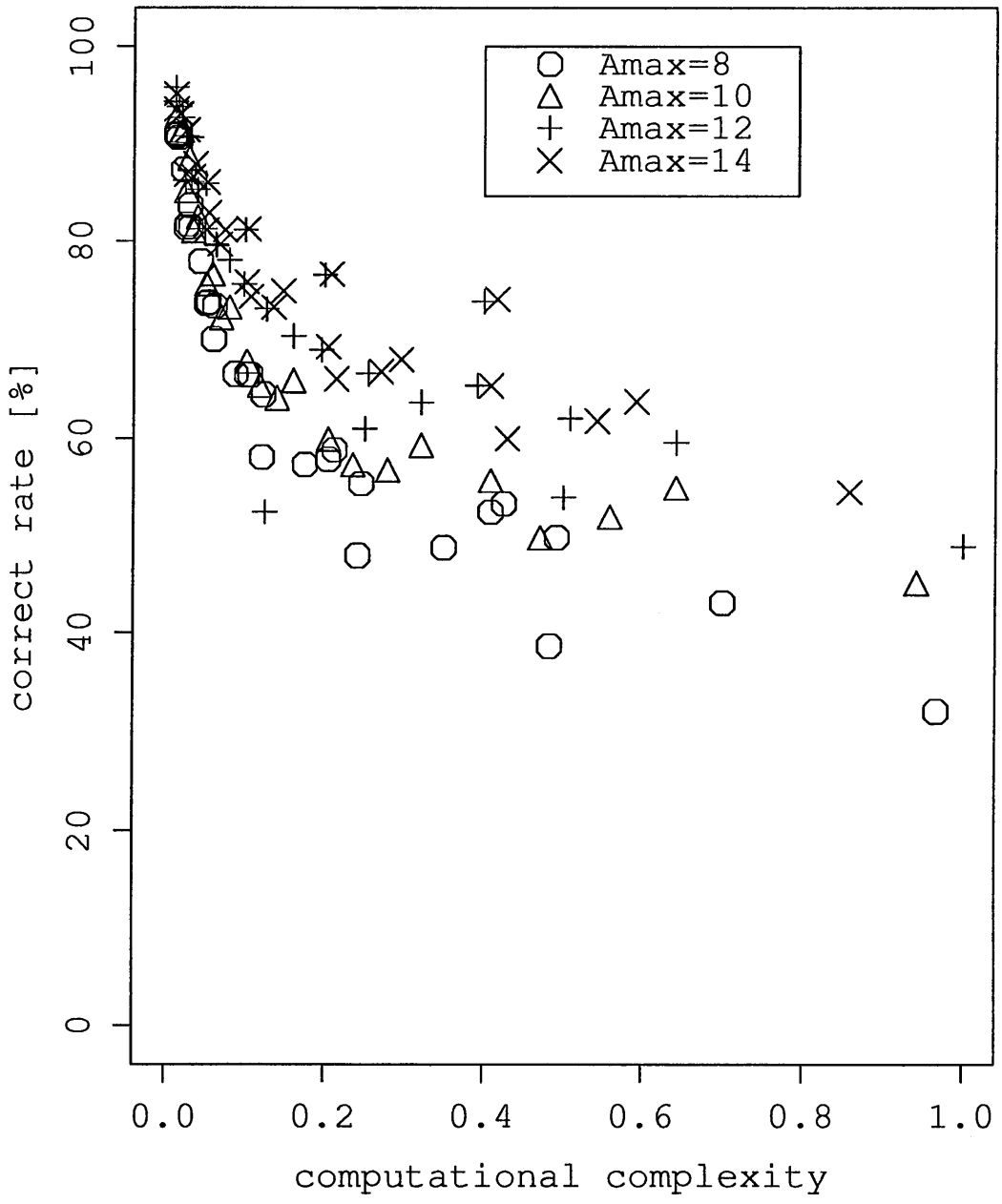


図 4.5　計算量と分類率 (複数符号)

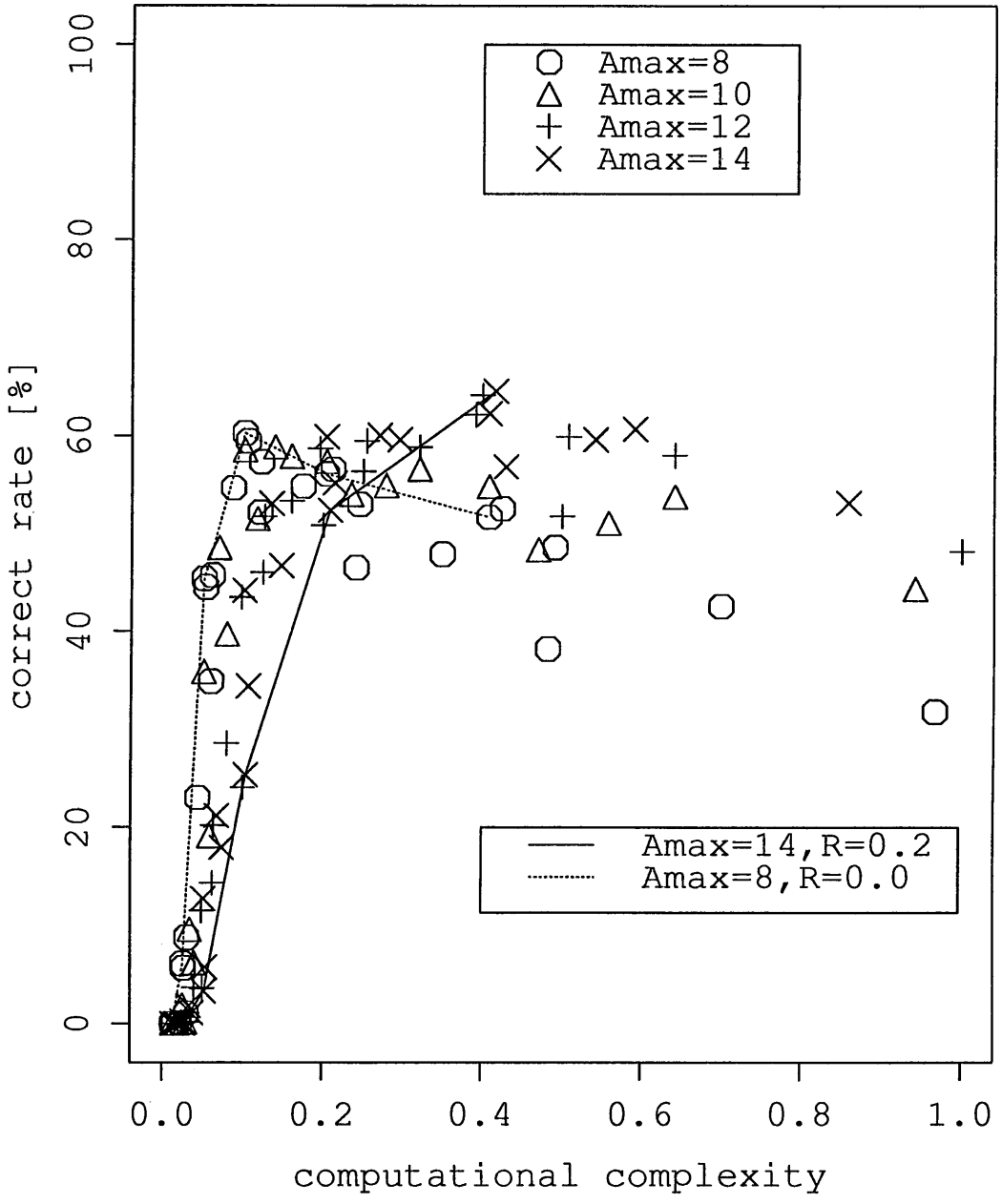


図 4.6　計算量と正解率 (複数符号)

がって，一つの部分領域だけによる大分類の実験をする必要があると思われる．また，一つの部分領域において，同一文字パターンの複数符号のクラス番号が異なった場合，その複数のクラスを一つのクラスとみなしてしまうことも有効かと思われる．つまり，クラスタリングの結果得られた複数の代表ベクトルのうち，似た様な代表ベクトルのクラスをまとめて同一クラスとしてしまうのである．しかし，まとめる方法も数多く考えられ，どの様なまとめ方が良いかを考察する必要がある．以上の様な方法でさらに符号の一致性を高めることが期待できる．

4.5　まとめ

本章では，最初に，部分領域の組合せを作成するときの問題点について述べ，それを考慮した組合せ法を基に部分領域の組合せを作成した．次に，部分領域の組合せから文字パターンを符号化する方法を述べた．そして，符号を基にした大分類の実験を行った．その結果，大分類には面積の大きい部分領域を少数だけ組合せたほうが有効であることが分った．さらに，より一致性の高い複数符号を提案し，単一符号に較べて大分類に有効であることを確認した．しかし，安定度が 1 となる部分領域が存在しないこと，複数の部分領域を組合せていること，により完全な大分類には至っていない．

第5章

結論

5.1 本研究のまとめ

本研究では，文字パターンは一定数の部品の組合せから成り，その組合せから文字パターンの符号化を行うことができるという考えを根本においた．そして，部分領域での文字パターンを部品とすることを前提とし，部分領域の選び方と部分領域の中でパターンをいかに，かつ，いくつに分類すべきかを明らかにすることを目的として研究を進めてきた．本研究で行ったことは以下の通りである．

- 部分領域における文字パターンの性質の調査
  文字パターンの部分領域を評価するため，複数の同一文字パターンが同じクラスに属する割合が高いほど大きな値をとる安定度を導入し，これにより，部分領域における文字パターンの性質を調べた．この果結，漢字の偏の存在する部分領域においては，安定度が高いことが分った．これは本研究で考える符号化には有効である．また，統計的に部分領域の評価をしたことで，これまでに分析されていなかった，どの様な部分領域が有用かを確かめることができた．

- 文字パターンの符号化
  文字認識に望まれる性能一つである処理の高速化のため，部分領域を基に文字を符号化する方法を示した．まず，部分領域の組合せを作る方法を提案し，実際に組合せを作った．文字を符号に変換して大分類を行う手法を提案し，実験を行った．しかし，良好な結果が得られなかったので，部分領域において同一文字パターンの属するクラスを複数にすることにより，標準符号と未知入力文字パターンの一致性を高めることができた．これにより，符号化による大分類の可能性を確認した．

5.2 今後の課題

本研究では完全な大分類には至っていないので，以下の点を改善する必要がある．

- 部分領域の面積の拡大
  面積が大きい部分領域ほど符号化に有効であることが分ったので，さらに面積に上限を設け実験を重ねる必要がある．また，今回は面積の下限を一定としたが，下限を変えることの影響についても検討する必要がある．

- クラス数の検討
  現在では全ての部分領域について一定のクラス数であるが，部分領域の面積に応じてクラス数を変化させることの影響を見る必要がある．面積の小さい部分領域に対してはクラス数が小さいほうが良好な結果が得られ，面積の大きい部分領域に対してはクラス数が大きいほうが良好な結果が得られている．したがって，部分領域の面積に応じてその領域で分類させるクラス数を変化させることを検討する必要がある．

- クラス割り当ての改善
  現在ではクラスの代表ベクトルとの整合から割り当てられるクラスを決定しているが，さらに同一文字の符号の一致性を高めるために，クラス割り当ての方法を検討する必要がある．一つの具体的な考え方は，複数符号の一つの部分領域におけるクラス番号が複数種となったときには，それらのクラスを同一のクラスとみなしてしまうことである．

これらを実現することで，本研究の目標である符号化に近づけることが可能となると思われる．

謝辞

本研究を進めるにあたり，全般的な御指導および御鞭撻を賜りました東北大学工学部通信工学科阿曽弘具教授に心より感謝いたします.

本研究をまとめるにあたり，貴重な御意見を頂きました東北大学大学院情報科学研究科丸岡章教授，同海老澤丕道教授に深く感謝いたします.

研究室内のゼミにおいては，東北大学工学部通信工学科成富敬助手，東北大学情報処理教育センター大町真一郎助手に多くの御助言を頂きました．ここに深く感謝いたします。

研究員として同じ研究室で研究を行い，普段から些細な質問にも親切にお答え下さった八戸高等専門学校細越淳一氏に深く感謝いたします.

研究方針や研究の進め方など細部にわたり御討論下さった阿曽研究室後藤英昭氏，同黒岩丈介氏に深く感謝いたします。

最後に，本研究を行うための計算機環境を常に良い状態に保っていただき，また，研究生活全般において大変お世話になった阿曽研究室の皆様に心よりお礼申し上げます.

学会発表

- 山田光影, 阿曽弘具 :" 文字パターンに基づく文字の符号化", 平成 6 年度電気関係学会東北支部連合大会 2D14.

参考文献

[1] 飯島泰蔵："パターン認識理論", 森北出版 (1989).

[2] 長谷博行, 米田政明, 酒井充, 吉田順作："手書き漢字の部分パターンの弾力的抽出方法", 信学論 (D) **J67-D**, 8, pp829–836(1984).

[3] 馬場口登, 相原恒博, 真田英彦, 手塚慶一："構造的セグメント整合による手書き漢字部分パターンの抽出と同定について", 信学論 (D) **J68-D**, 3, pp337–344(1985).

[4] 江島俊朗, 勝山裕, 木村正行："特徴ベクトルの分割と統合による手書き文字の大分類", 信学論 (D) **J70-D**, 2, pp.398–404(1987).

[5] 孫寧、安部正人、根本義章："部分パターンを用いた手書き文字の詳細識別", 平成5年度電気関係学会東北支部連合大会講演論文集, p141(1993).

[6] 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", 信学論 (D-II), **J74-D-II**, 3, pp.330–339(1991).

[7] C.J.Hilditch: "Linear skelton from square cuboards", In Machine Intelligence 6, B.Meltzer & D.Michie, Eds, Univ. Press, Edinburgh, pp.403–420 (1969).

[8] Linde, Y., Buzo, A. and Gray, R.M.: "An Algorithm for Vector Quantizer Design", *IEEE Trans. Comm.* Vol. **COM-28**, No.1, pp84–95(1980).