文字構造情報に基づく高精度な文字切出し処理を用いた文書認識システム†*

孫 寧†† 鈴木雅人††† 根元義章††
阿曽弘具†††† 木村正行†††††

文書認識システムにおいて，文字切出しの精度の向上は大変重要な課題である．現状では，個別文字の認識精度に比べ，切出しの精度が幾分低いため，文書認識システム全体の性能は切出しによって大きく左右されることになる．切出しの精度に影響する要因は多数存在するが，新聞や文庫本のような文字中心のものに限って言えば，可変ピッチ文字列における分離文字や半角文字（縦書きの場合，半ピッチ文字）の存在が大きな要因である．すなわち，分離文字同士，あるいは分離文字と半角文字が二つ以上続いた場合に，その切出しが困難になる．本論文では，そのような場合の問題の解決策として，文字構造情報を用いた統合方法を提案する．この方法で，高速かつ正確な切出しが可能であることを示す．新聞社説（20部，約45,000文字）を対象に本手法を適用した認識実験を行い，文書中に分離文字や半角文字が連続して二つ以上続いた場合でも，99.12%の統合率が確保でき，また，文書画像の入力から認識結果の出力までのシステム全体として99.0%以上の認識率が得られることを示す．

1. はじめに

近年，文字認識についての研究がこれまでの個別文字中心のものから一般文書を対象とするものに変わりつつある．代表的なものとして，新聞，文庫本，帳票，辞書などがある[1]．文書画像を認識するシステムとして，まず行わなければならないことは個々の文字の切出しである．文字切出しは二つの段階に分けることができる．第一段階は，文書画像内の図形，罫線，飾り文字などの文字以外の領域と文字列からなるテキスト領域とを分離する，言わば，領域の分割処理である．第二段階では，分割された領域に対して，それぞれ個別処理を行う．新聞，文庫本のような文字中心の文書においては，テキスト領域の文字に対する切出しを行う．本論文では，この第二段階の切出しを主な対象とする．

現在，テキスト領域における切出しは，文字の大きさが一定で，かつ比較的隙間の空いている文字列に対し，ほぼ100%近い精度に達している反面，大きさの違う文字が混在し，不定ピッチで文字が並んでいる文字列に対しては，依然切出し率が低迷している[2),3)]．その一つの理由は，黒画素の塊単位で文字候補図形を切り出し，それをうまく統合して文字図形を得る方法では，文字が黒画素の一つの塊とはなっていない分離文字があるとき，その一部分を抽出してしまい，それと文字列中に存在する大きさの小さい文字（半角，半ピッチ文字など）との区別が困難で，うまく統合できなかったことである．本論文では，この困難を克服するため，新しく文字構造情報を定義し，それを用いた高速で高精度な統合アルゴリズムを提案する．

文字の高精度切出しを目指して，これまで数多くの研究が行われてきた．それらにおける手法は，イメージ情報（文字の大きさやピッチ情報など）を利用した統合方法[4)]と認識を併用する統合方法[5)~8)]および単語情報を用いる方法[9),10)]に大別できる．単語情報を用いる方法は，未知単語に対する誤読の存在，複合語に対する分割処理，さらにべた書き文字列に対する文節認定など多くの難問を抱えており，切出しのために，より正確な認識が必要であるという自己矛盾もはらんでいる．イメージ情報を利用した方法は書式が簡単な文書を対象として開発されており，いろいろなサイズの文字を含む文書には適用しても間違うことが多い．認識を併用する手法は，アルゴリズム上複雑な操作が必要となるが，精度が比較的高い．この手法は分離文字，半角文字混じり文書に対する処理方法によって，さらに二つに分けられる．

---

† Document Recognition System Using High Accuracy Segmentation Algorithm Based on Information of Character Structure by NING SUN (Computer Center, Tohoku University), MASATO SUZUKI (Research Institute of Electrical Communication, Tohoku University), YOSHIAKI NEMOTO (Computer Center, Tohoku University), HIROTOMO ASO (Department of Information Engineering, Faculty of Engineering, Tohoku University) and MASAYUKI KIMURA (Japan Advanced Institute of Science and Technology, East).

†† 東北大学大型計算機センター
††† 東北大学電気通信研究所
†††† 東北大学工学部情報工学科
††††† 北陸先端科学技術大学院大学

* 本論文の概略は文献20)として発表されている．

- **組合せ法** この方法はまず，候補図形に対し，物理的な大きさによって，統合可能となるすべての組合せを作る．次にこれらの組合せに対してマッチングを行い，文字列として最も距離の小さいものを選ぶ手法である[5),6)]．
- **部分パターン法** この方法は分離文字に対し，あらかじめその部分パターンも普通の文字パターンと一緒に辞書に登録する．さらに，これらの部分パターンに関する接続情報もテーブルとして用意する．例えば，“北”のような分離文字の場合，左の部分パターンと右の部分パターンを辞書にそれぞれ登録し，さらに，用意した接続テーブルに左の部分パターンに対し，右の部分パターンへの接続許可を登録する．判定は認識の結果および接続テーブルを用いて行う[7),8)]．

組合せ法と部分パターン法の有効性は認識実験によって確認されている．しかし，別な観点からこの二つの方法にはそれぞれ欠点がある．組合せ法では，候補図形に対して多くの組合せをつくらなければならないことから，処理時間がかかるという点である．この傾向は特に統合候補図形の数が多くなった場合顕著に現れる．一方，部分パターン法では，文字の部分パターンの登録や接続テーブルの作成などの作業の自動化が困難なため，システムの構築に多くの労力が必要であるという点である．自動化が困難な理由の一つに接続対象が複数に渡る可能性があることが挙げられる．

本論文で提案する文字切出し手法はこれらの欠点を克服し，高速，高精度でありながら，判定用文字構造辞書の作成が簡単であるという特徴を持っている．本手法は，候補図形の数によらず，常に候補図形列の先頭から順に判定を行うことにより，組合せ法のようにすべての組合せを作らなくても済み，高速性を実現する．また，統合判定を行う際，3種類の文字構造情報（分離情報，線分情報，線素情報）を用いて，より正確な判定を可能にし，高精度化を実現する．さらに，本手法における統合用の文字構造辞書の作成において，3種類の文字構造情報の抽出をそれぞれ数十ステップのプログラム（C言語）で実現し，多数の文字に対しても作成が自動化され，容易になっている．これは文字単位内での構造情報の抽出が文字間のそれ（部分パターン法）と比べると，はるかに簡単であることによる．

本研究では，新聞社説を対象として，文字切出しアルゴリズムおよび全体システムの性能を実験的に評価する．新聞文字を対象にした認識実験として，これまで数例[11)~13)]が報告されている．これらの実験は，当時の計算機の性能に制限されたためもあり，いずれも社説数部（あるいは数千字）程度を対象とし，切出しを含めたシステム全体の実用的性能は未知のまま残されていた．本研究では，新聞社説20部(約45,000字)を対象に認識実験を行い，実用的性能を明らかにする．すなわち，分離文字同士，あるいは分離文字と半ピッチ文字が続いた454例のうち，450例について正しく切り出すことができ，99.12%の統合正解率が得られることを示す．また，システム全体として，99.0%以上の認識率が確保できることを示す．なお，本研究において，文字パターンの特徴量として活字および手書き文字認識において有効性が確認されている方向線素特徴量[14),15)]を用いた．方向線素特徴量は手書き文字を対象とした方向パターンマッチング[16)]や加重方向指数ヒストグラム法[17)]における特徴量に類似しているが，特徴抽出処理がより簡潔で，活字文字に対しての性能も明らかになっている．

本論文の構成は，まず，第2章でシステムの基本構成および認識対象文書に合った標準パターン辞書作成法について述べたあと，第3章で，本研究で提案する統合アルゴリズムについて述べる．第4章では，行った認識実験について結果を示し，考察を行う．第5章で，結論をまとめる．

2. 文書認識システム

本研究では，実験用文書データとして，読売新聞の社説を1989年4月1日から70日間に渡り，収集した．この70部のうち，50部を辞書作成用に20部を未知入力用に使用する．実験用計算機はSUN 4/110を使用し，スキャナとして読み取り密度が10本/mmのものを使用する．

2.1 基本構成

文書認識システムの基本構成を図1に示す．このシステムはデータ2値化，傾き補正，領域分割，文字候補図形の検出，イメージ情報による切出し，前処理，特徴抽出，認識・統合処理によって構成されている．この中で，文字構造情報を用いた統合処理が本システムで新たに提案するものである．以下，システムの主な部分について概要を述べる．

文字候補図形の検出 文字候補図形とは黒画素の塊のことであり，文字，文字の一部，複数の文字（接触文字の場合）となるもので，以下では単に，候補図形

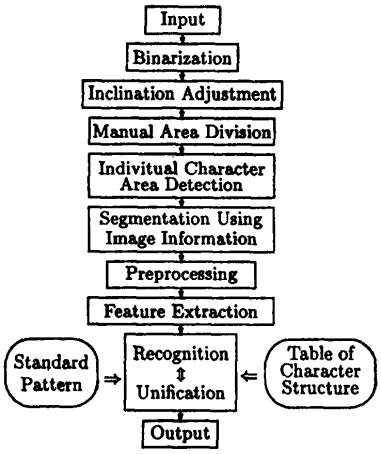


図 1 文書認識システムの概略図
Fig. 1 Document recognition system.

とも呼ぶ．候補図形は黒画素の接続関係を調べるラベリング処理で検出する．

候補図形検出後の状況を図2に示す．図2では，確認できるように候補図形ごとに長方形で囲んである．候補図形は黒画素の連結した部分を囲む長方形として切り出されるため，非分離文字（“た”なども含む）は一つの図形として得られるが，仮名の“ハ”，漢字の“二”，“三”のような文字は二つ以上に分かれた候補図形として得られることになる．

イメージ情報による切出し 縦書きのテキスト領域に対して，イメージ情報，すなわち，文字図形の平均サイズに基づく切出しを以下の手順で行う．

［イメージ情報による切出しアルゴリズム］

1. 文字の平均幅および平均高さの推定 候補図形のサイズで最も頻度の高いものからいくつかを選んで，その幅と高さの平均を求めて，文字平均幅および文字平均高さとする．
2. 列に関する情報の推定 検出された候補図形を横軸に射影し，その分布から列の同定をし，各列の幅，ピッチを決定する．
3. ノイズ除去 2×2ドット以下の孤立した黒画素を消去する．
4. 文字候補図形のラベリング 候補図形の中心座標と同定された列の位置情報とから候補図形がどの列に属しているかを判定し，属する列の番号でラベルづけする．
5. 列内での統合処理 以下の条件がすべて満たされる候補図形を統合する．
   - 同一列内にある．
   - 縦軸への射影が重なりを持っている．
   - 統合後の図形の高さが文字平均高さの1.1倍を超えない．

   （これにより，“北”，“川”のような水平方向に分離している文字が統合されることになる．）
6. 接触文字の分離処理 候補図形の高さが平均文字高さの1.1倍を超えたものを接触文字と判定し，分離処理を行う．候補図形の縦方向の先頭から文字平均高さの整数倍の位置を基準点とし，上下±10ドット以内で縦軸への射影を取り，その度数の一番低いところを分離点として分離する．分離された図形で種類“4”を割り当てる．
7. 非分離文字の優先切出し 平均幅，平均高さに近いサイズの候補図形について，上下の図形と統合を行うと，統合した図形の高さが文字平均高さの1.1倍を超えてしまうとき，非分離文字として切り出す．切り出された図形に種類“0”を割り当てる．
8. 記号類の抽出 “，”，“。”，“・”のような小さい記号を切り出す．記号であるかどうかの判定は候補図形の大きさおよび行内の存在位置によって行う．判定された図形に種類“1”を割り当てる．
9. 強制統合 平均文字サイズに基づいて，次の条件がすべて満たされる候補図形を統合する．その結果に“2”を割り当てる．
   - 統合候補図形は同一列内にある．
   - 統合（複数図形可）しても，文字平均高さの1.1倍を超えない．

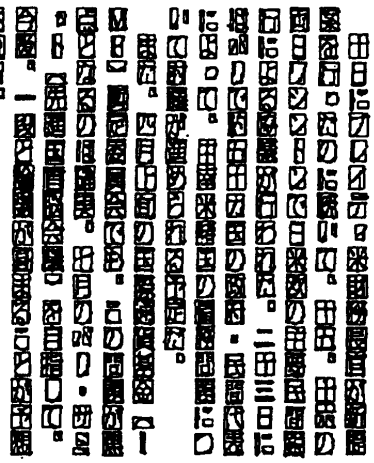


図 2 候補図形検出後のイメージ
Fig. 2 Image of character candidates.

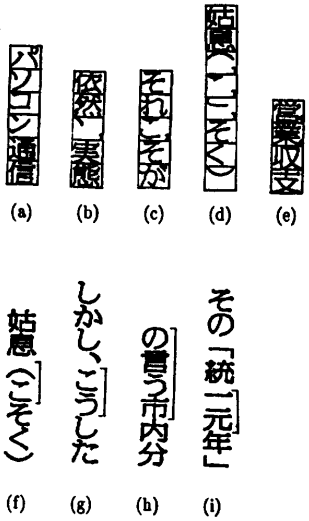


図 3 イメージ情報による切出しの結果
Fig. 3 Results of segmentation by image information.

- 統合対象となる候補図形の上下は非分離文字である.
- 記号類ではない.

(これによって，縦文字列「十三人」のような，分離文字の上下に非分離文字が存在する場合，統合が行われる.)

10. 保留図形 以上の操作で種類が割り当てられなかった図形に種類“3”を割り当てる．保留図形という.

図3はイメージ情報による切出し後の様子を示している．(a)から(e)までは候補図形の取り得る五つの種類0～4を枠の右下隅に示している．ただし，種類“0”は無印としている．種類“3”の保留図形は(f)～(i)のような分離文字同士，あるいは分離文字と半ピッチ文字が連続した文字列である（図において保留となったものを“]”で表している)．保留図形は単なるイメージ情報だけでは正確な切出しが困難であり，本論文で提案する統合アルゴリズムの対象となる.

前処理 前処置はスムージング，正規化，細線化，線素化の四つの過程からなる.

特徴抽出 線素化データから，方向線素特徴量を求める.

認識および統合処理 切出し済みの文字に対して標準パターン辞書を用いて認識を行い，保留図形に対して文字構造情報を用いた統合処理を行う．標準パターン辞書の作成については次節で，統合処理については第3章で詳しく述べる.

2.2 標準パターン辞書の作成

一般的に，標準パターンのフォントが未知入力のそれに近ければ近いほど高い認識率が期待できる．ここでは，収集した新聞社説70部のうち，50部を用いて各文字フォントを収集し，標準パターン辞書を作成することにした．収集の迅速化をはかるため，初期標準パターン（字種数 3,303）として，レーザプリンタのフォントを用い，標準パターン作成用の社説データを認識させながら，その結果を人間が確認し，標準パターンを新聞の文字フォントに置き換える方法を用いた．収集にあたり，各字種につき10サンプル程度を目標とし，最終的な標準パターンは集まったサンプルの平均を用いた．集まらなかった字種はそのままレーザプリンタのフォントで代用する.

この辞書作成の効果について，**表1**，**2**に実験結果を示す．表1は，レーザプリンタのフォントを標準パターンとし，未知入力用の社説10部（正しく切り出された約 22,000 文字）を認識した結果である．1st，2nd，10th はそれぞれ1位認識率，2位累積認

表 1 LASER SHOT のフォントによる実験結果
Table 1 Result of recognition using LASER SHOT font.

セット	1st	2nd	10th
No. 1	87.06	93.03	98.34
No. 2	86.08	94.14	98.85
No. 3	85.71	93.79	98.52
No. 4	87.39	93.90	98.54
No. 5	84.25	90.23	94.13
No. 6	88.84	94.32	99.05
No. 7	84.65	92.09	98.87
No. 8	86.07	93.52	99.18
No. 9	85.29	93.66	98.31
No. 10	85.15	95.21	99.08
平 均	86.04	93.39	98.29

表 2 収集字種数と認識率の関係
Table 2 Relation between characters and recognition rate.

セット数	収集字種数	1位認識率	2位認識率	10位認識率
10	1,220	97.28	98.95	99.36
30	1,578	98.39	98.73	99.51
50	1,705	98.84	99.21	99.21

識率，10 位累積認識率を表す．1位の平均認識率が 86.04% に留まったが，人間による確認作業の効率化には役立つことを示している．誤認識が多かった理由は図 4 で示すようにフォントが違うためと思われる．

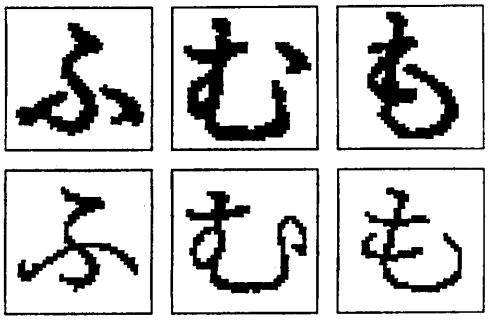


図 4 フォント形状の違い
(上段：レーザプリンタ，下段：新聞)
Fig. 4 Differences between fonts.
(upper : Laser printer, lower : newspaper.)

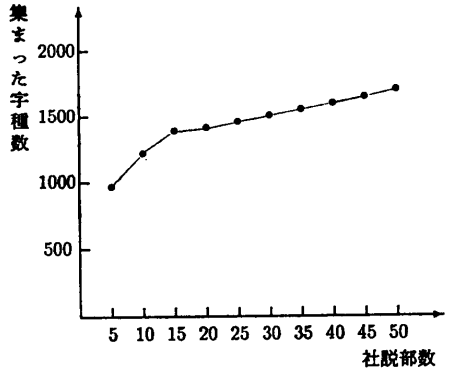


図 5 社説部数と収集文字数の関係
Fig. 5 Relation of the number of editorials and the number of category of characters.

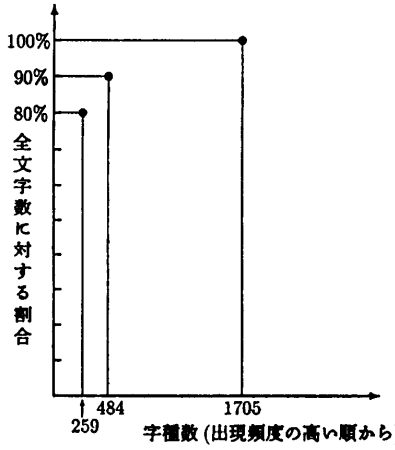


図 6 文字の出現頻度
Fig. 6 Frequency of occurrences of characters.

特に平板名の“ふ”“ぶ”“ぷ”のような互いに微妙に違う文字の間でかなりの誤りが生じていた．

表 2 は，50 部の辞書作成用データから辞書を作成しつつ行った，未知入力用の社説 5 部（約 11,500 文字）についての認識実験結果である．この結果は，収集字種数（標準パターンを置き換えた字種数）が増加するにつれ，認識率が上昇することを示しており，標準パターンの置き換えにより 10 部程度の収集でも認識率が 11% も上がることがわかる．認識率向上の一つの理由は字種の出現頻度にあると考えられる．実際に，辞書作成用の社説 50 部（約 114,724 文字）について字種の出現頻度を調査したところ次のようであった．対象部数の増加により，収集できた字種数が増加する様子を図 5 に，収集した各字種の割合いを大きい順に図 6 に示す．図 6 からは，文字パターン全体の 90% が，実際に現れた字種の 1/3 以下の字種からなっていることがわかる．すなわち，この高頻度の字種に対する認識が正しく行われれば，認識率が 90% 以上になることが保証される．なお，新聞一年間分についての頻度調査が文献18)において発表されているが，図 6 の結果が，その一般的な傾向を反映していることが確認された．

3. 文字構造情報を用いた統合アルゴリズム

統合処理では分離文字，半ピッチ文字が存在する候補図形に対し，各文字の境界を見つけ，切り出す操作を行う．ここでは，その対象は図 3 の(f)～(i)のような保留された候補図形である．

3.1 文字構造辞書の作成法

認識対象のすべての文字に対して，文字ごとに文字の構造を反映する以下の 3 種類の情報を求め，文字構造辞書を作成する．

1. 分離情報：垂直方向の分離数に関する情報
   - 4 dot 以上分離している箇所の数 D_4
   - 16 dot 以上分離している箇所の数 D_{16}
2. 線分情報：水平方向の線分数に関する情報
   - 長さが 8 dot 以上の線分数 L_8
   - 長さが 32 dot 以上の線分の数 L_{32}
3. 線素情報：4 種類の方向線素数に関する情報
   - 水平方向の線素の数 E_1
   - 垂直方向の線素の数 E_2
   - $+45°$ 方向の線素の数 E_3
   - $-45°$ 方向の線素の数 E_4

垂直方向の分離数および水平方向の線分数は細線化後

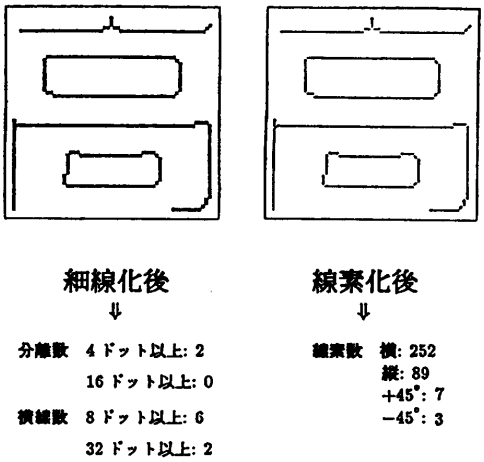


図 7 文字構造情報の例
Fig. 7 Example of character structure information.

に，4種類の線素数は線素化後にそれぞれ求める．図7に例として漢字“高”に対する文字構造辞書の内容を示す．

辞書の3種類の情報は次の意図で選定した．スキャナの読み取り誤差やノイズなどによる影響で，未知候補図形の分離数や線分数は変化する．その変化に対処するため，それぞれの数の上・下限を与えるものとして，二つの数を用意した．また，かすれなどによる影響を排除するために，3 dot 以下の分離を無視している．縦書き文書における分離文字の多くは水平方向に長い線分を持っているため，その構造を反映する簡単な情報として，水平方向の線分だけを計数している(当然横書き文書の場合には垂直方向の線分を計数する)．この文字構造辞書を用いることによって，候補図形の分離情報，線分情報，線素情報を相互補完的にチェックし，正確な判定を行うことを意図している．

3.2 統合処理の流れ

統合アルゴリズムを図8に示す．図8の N_0 は保留した候補図形の個数で，M_0 は N_0 個の図形の先頭から合併可能な最大個数である．すなわち，M_0 は合併した場合の図形の高さが文字平均高さの1.1倍以内(予備実験により選んだ)となるように決める．統合判定は仮の文字図形 L と，その図形を認識し得られた1位候補文字との間で文字構造情報を照合することで行う．

仮の候補図形 L の8ドット以上の分離数を d_8，16ドット以上の線分数を l_{16}，縦，横，+45°，−45°の線素数をそれぞれ e_1, e_2, e_3, e_4 とし，照合は以下の条件がすべて満たされる場合に，統合可と判定する．

垂直方向の分離数に関する条件：$D_4 \geq d_8 \geq D_{16}$

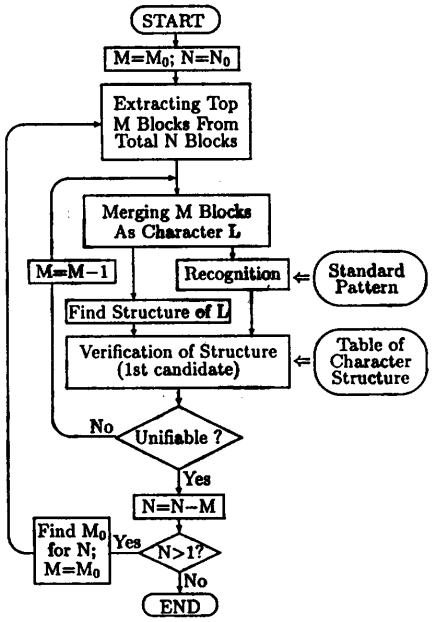


図 8 文字構造情報を用いた統合アルゴリズム
Fig. 8 Unification algorithm using character structural information.

ただし，D_4, D_{16} は L の1位候補の図形の 4 dot および 16 dot 以上の分離数である．

水平方向の線分数に関する条件：$L_8 \geq l_{16} \geq L_{32}$

ただし，L_8, L_{32} は L の1位候補の 8 dot および 32 dot 以上の線分数である．

4種類の方向線素に関する条件：

- $|e_i - E_i| \leq C_1 \quad i=0 \sim 3$
- $\sum_{i=0}^{3} |e_i - E_i| \leq C_2$

ただし，E_1, E_2, E_3, E_4 は L の1位候補の対応する線素数である．また，C_1, C_2 は実験によって決める定数で，予備実験により，C_1 を 30，C_2 を 60 とした．

統合アルゴリズムの適用について，図9を用いて説明する．図9は候補図形 L が現に存在しない文字図形の場合である．統合候補図形の総数 N_0 は3で，先頭から合併できる最大ブロック数 M_0 は2である．したがって，候補図形 L は二つの図形からなり，図に示すような図形となる．これを認識した結果，“2”という1位候補が得られる．この結果を用いて，辞書内の“2”に関する情報と図形 L の構造情報とを照合すると，分離数に関する条件が満たされていない．よって，統合不可となる．次に $M=1$ として処理が続き，結果的に単独の括弧が切り出されることになる．

以上のように，この統合処理において，仮の図形 L

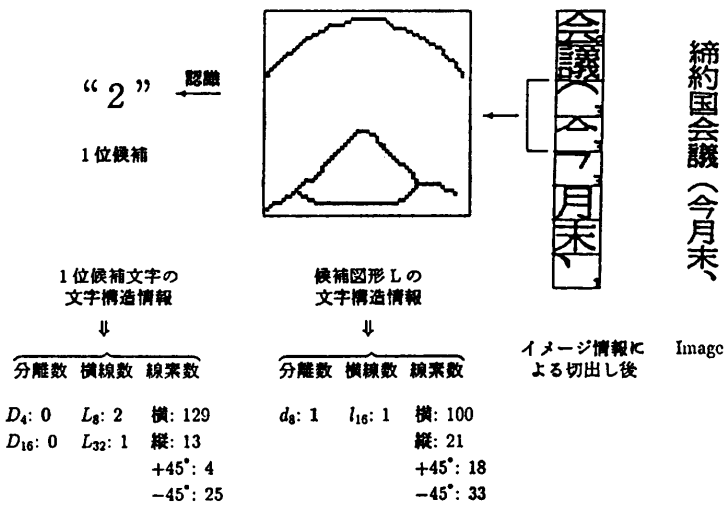


図 9 統合処理の例
Fig. 9 Example of unification.

が現に存在する文字図形の場合，L を認識し，得られた1位候補が正しいならば，統合可となることが期待される．また，この図形 L が現に存在しない文字図形の場合，認識によって図形 L が標準パターン内のどれかの文字であると認識されるが，照合によって，統合不可と判定されることが期待される．これらが期待どおりになるかどうかを実験によって確かめた．

4. 実験結果および考察

認識実験は2.1節で述べた文書認識システムを用いて，未知入力 20 部（約 45,000 文字）について行った．

4.1 統合処理

イメージ情報による切出しで文字パターンの約98%が切り出された．残りが保留図形として，統合処理の入力となる．保留となった分離文字同士，あるいは分離文字と半ピッチ文字が連続した候補図形の系列が454例存在し，それらに統合アルゴリズムを適用した結果，450例が正しく統合された．すなわち，99.12%の統合正解率が得られた．

イメージ情報および構造情報による統合の失敗例は次のとおりである．

1. イメージ情報による統合の際，"へ" と "一" が隣合った場合，この2文字を統合すると，高さが文字平均高さ以下となってしまうため，イメージ情報による切出しの段階で，一つの文字として，強制統合されてしまった．
2. 構造情報による統合の失敗例として，図3の(i)のようなものがある．この例の場合，"一" と "元" の上の部分を合わせた図形が漢字の "二" に認識され，分離数，線分数，線素数のどれを取っても，不一致が検出されなかったため，誤って統合を行ってしまった．また，"元" の下の部分は "凡" と認識され，統合の誤りを検出することはできなかった．

このような統合の誤りを解決するには，最終的に単語や文章レベルでの判断が必要であると思われる．たとえば，上の 2. の場合，認識結果が "二凡" となるが，そのような単語は存在せず，誤りであるという判定が可能になる．このように文章レベルで意味のある文章として成立するかどうかを判断し，その結果をフィードバックして使えば，より正確な統合処理を行うことができると思われる．このような研究は単語情報による統合処理とも関連して，独立した課題として，今後の発展が望まれる．

以上のような統合時の問題点のほか，切出し時の失敗もある．すなわち，分離文字の一部がほかの文字と接触し，分離作業が失敗したケースである．このような問題を解決するには，切出しでの例外処理を組み込む必要があると考えられる．

4.2 文書認識

システム全体としての認識率を**図10**に示す．図10はセット別の認識率を1位認識率の高い順にソートして示している．この1位認識率を平均すると，99.0%

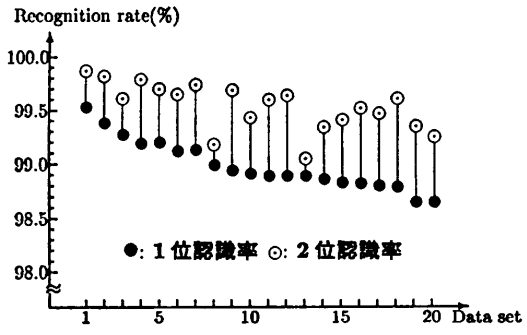


図 10 新聞社説20部（約45,000文字）についての認識結果
Fig. 10 Recognition rate of characters in 20 editorials.

以上となる．2位までで は 99.53% が得られた．

一部の誤認識は “ぱ” ⇔ “ば，“大” ⇔ “太”，“土” ⇔ “士” のような類似文字間で起きている．今回は，統合処理に焦点をあてたため，組み込んでいないが，類似文字識別法により，解決可能な誤認識である．

4.3 処理時間

イメージデータの取り込みのあとの切出しから結果出力までの処理は同一ワークステーション上で行い，一文字についての認識時間は 0.85 秒であった．この中で，最も時間がかかった処理は前処理 (0.26 秒) と認識 (0.51 秒) であった．

5. おわりに

本論文では，分離文字，半ピッチ文字を含む不定ピッチ文字列に対し，文字構造情報を用いた統合アルゴリズムを提案した．提案したアルゴリズムは

- アルゴリズムが簡単
- 統合所要時間が短い
- 構造辞書の作成が簡単
- 統合正解率が高い

などの長所があり，新聞社説を対象に実験を行った結果，99.12% の統合正解率が得られた．また，この統合アルゴリズムを用いた認識システムで，99.0% 以上の認識率が得られた．よって，本論文で提案した統合アルゴリズムおよび認識システムは活字印刷物の認識において，有効であると判断される．

この認識アルゴリズムの適用のために，認識対象の文書から文字フォントを抽出し，標準パターン辞書を半自動的に作成する方法を提案した．高精度な認識が達成できた一つの理由がこの認識対象文書に合った辞書を用いたことにあることも実験データとして示した．今後，より効率的な辞書作成法が望まれる．

新聞における文字の出現頻度の偏りについても，実験対象データについて確認した．一般的に，頻度の高い文字についての高精度認識を実現することにより，文書全体の認識率を上げることができる．例えば，高頻度の類似文字 (“ぱ” ⇔ “ば”，“大” ⇔ “太” など) 間の誤認識を少なくするため，類似文字識別法[19]などを本システムに組み込むことで，認識率をさらに高めることができる．それは，今後の課題である．

謝辞 本研究を行うにあたり，ご討論いただいた堀口 進助教授，下平 博助手，阿部 亨助手をはじめ，木村研究室の方々に感謝します．また，本研究の一部は文部省科学研究費特別推進研究 (No. 63060001) の援助を受けた．

参考文献

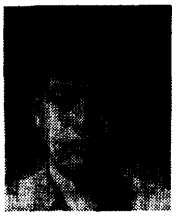
1) 坂井邦夫：文字・文書の認識と理解，信学誌，Vol. 71, No. 11, pp. 1182-1191 (1988).
2) 電子情報通信学会：パターン認識，電子情報通信学会 (1988).
3) 飯島泰蔵：パターン認識理論，森北出版(株) (1989).
4) 佐藤道弘，木田博巳：不定ピッチ文字列を含む印刷文書における文字切出し手法，信学技法，PRU 88-159 (1988).
5) 村瀬 洋，若原 徹，梅田三千雄：候補文字ラティス法による枠無し筆記文字列のオンライン認識，信学論(D)，Vol. J68-D, No. 4, pp. 765-772 (1985).
6) 鈴木昭浩，金井 浩，川添良幸，牧野正三，城戸健一：切り出しと認識を同時に行なう活字デーヴァナーガリ文献の認識法，信学論 (D-II)，Vol. J72-D-II，No. 10, pp. 1643-1649 (1989).
7) 古田雅裕，中村憲司，木下修逸，木村正行：認識を用いた自動文字切り出し，1988 信学春季全大，D-474 (1988).
8) 宮原末治，木村義政，豊田 充，宮田一人：部分パターンによる可変ピッチ文書からの文字切り出しと認識，信学論(D)，Vol. J72-D-II，No. 6, pp. 846-854 (1989).
9) 村瀬 洋，新谷幹夫，若原 徹，小高和己：言語情報を利用した手書き文字列からの文字切り出しと認識，信学論(D)，Vol. J69-D, No. 9, pp. 1292-1301 (1986).
10) 佐藤慎治，辻 善丈，津雲 淳：制限付文字列読み取りの一検討，信学技法，PRU 88-115 (1988).
11) 西村 康，野口要治，豊田順一：新聞記事の本文を構成する文字の切出し，第 24 回情報処理学会全国大会論文集，3E-7 (1982).
12) 豊田順一，野口要治，西村 康：日本語印刷文書における文字の切出し，情報処理学会論文誌，Vol. 24, No. 4, pp. 481-487 (1983).
13) 秋山照雄，内藤誠一郎，増田 功：非接触文字優先切出しによる印刷物からの文字切出し法，信学論(D)，Vol. J67-D, No. 10, pp. 1194-1201 (1984).
14) 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)，Vol. J74-D-II，No. 3, pp. 330-339 (1991).
15) 孫 寧，曽阿弘具，木村正行：連想整合法に基づく高速文字認識アルゴリズム，情報処理学会論文誌，Vol. 32, No. 3, pp. 404-413 (1991).
16) 斉藤泰一，山田博三，山本和彦：手書き漢字の方向パターン・マッチング法による解析，信学論(D)，Vol. J65-D, No. 5, pp. 550-557 (1982).
17) 鶴岡信治，栗田昌徳，原田智夫，木村文隆，三宅康二：加重方向指数ヒストグラム法による手書き

漢字・ひらがな認識，信学論(D)，Vol. J 70-D, No. 7, pp. 1390-1397 (1987).
18) 現代新聞の漢字，国立国語研究所報告，Vol. 56，秀英出版(株) (1976).
19) Sun, N., Uchiyama, Y., Ichimura, H., Aso, H. and Kimura, M.: Intelligent Recognition of Characters Using Associative Matching Technique, *Proc. Pacific Rim Int. Conf. on Artificial Intelligence '90*, pp. 546-551 (1990).
20) 孫　寧，鈴木雅人，阿曽弘具，木村正行：文字の分離及び構造情報を用いた新聞切り出しシステム，1990 信学春季全大，D-503 (1990).

(平成3年8月26日受付)
(平成4年7月10日採録)

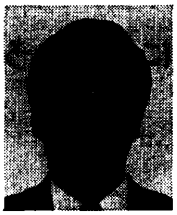
孫　　寧（正会員）

昭和37年生．昭和61年職業訓練大学校電子科卒業．昭和63年東北大学大学院工学研究科情報工学専攻修士課程修了．平成3年同大大学院博士課程修了．同年同大大型計算機センター助手．工学博士．文字認識，コンピュータネットワークなどの研究に従事．電子情報通信学会，人工知能学会各会員．



鈴木　雅人

昭和43年生．平成2年東北大学工学部電気・情報系学科卒業．平成4年同大学大学院工学研究科情報工学専攻修士課程修了．現在，同大学大学院博士課程在学中．文字認識，プログラミング言語の意味論に関する研究に従事．



根元　義章（正会員）

昭和20年生．昭和43年東北大学工学部通信工学科卒業．昭和48年同大学院工学研究科博士課程修了．同年同大工学部助手．昭和59年同大電気通信研究所助教授．平成3年同大大型計算機センター教授，現在に至る．工学博士．マイクロ波伝送回路，衛星利用ネットワーク，情報伝送システム，画像処理の研究に従事．昭和56年IEEE, Microwave Prize (MTT-S, 論文賞) 受賞，IEEE，電子情報通信学会各会員．

阿曽　弘具（正会員）

昭和21年生．昭和43年東北大学工学部電気卒業．昭和48年同大大学院博士課程修了．同年同大工学部助手．昭和54年名古屋大学工学部講師，昭和54年同助教授，昭和61年東北大学工学部助教授を経て，現在，同教授，工学博士．その間，学習オートマトン，セル構造オートマトン，並列処理理論，シストリックアルゴリズム設計論，文字認識などの研究に従事．昭和53年度電子通信学会学術奨励賞，平成3年度電子情報通信学会業績賞受賞．IEEE, ACM, EATCS, 電子情報通信学会，人工知能学会，LA 各会員．

木村　正行（正会員）

昭和2年生．昭和29年東北大学工学部電気卒業．昭和34年同大大学院博士課程修了．同年同大電気通信研究所助手．昭和37年同研究所助教授．昭和45年同大工学部教授．平成3年東北大退官．同年北陸先端大教授．図書館長．工学博士．システム理論とその応用，しきい値理論，学習オートマトン，視覚神経系のモデル(神経回路網)などの研究に従事してきた．最近，文字，図形および顔画像の認識，音声認識など知的情報処理の分野に興味を持っている．平成3年度電子情報通信学会業績賞受賞．