Journal of the Faculty of Engineering, Shinshu University, No.74,1994
信州大学工学部紀要 第74号


# 日本語文書読み取りシステムの試作

福田 亮治* 岡本 正行**

# An Experimental Implementation of a Japanese Document Recognition System

Ryouji FUKUDA* Masayuki OKAMOTO**

This paper describes the current state of developing a document recognition system. The principal stages of document recognition process are (1) image digitization and if necessary, skew detection and correction, (2) page segmentation which divide a document page into some areas such as text, line drawings or pictures, tables, and so on, (3) character segmentation and (4) character recognition (OCR). In our system, for the stages (1), (2) and (3), new algorithms have been developed and tested. For the stage (4), We have evaluated some methods for OCR which have been proposed, and found the best combination of those method in order to get good performance of OCR and efficient processing. This experimental system is totally implemented by software on the work station with X-window system. Some experimental results on printed documents shows the effectiveness of our system.

## 1. はじめに

近年，様々な書式を持つ印刷文書を自動的に計算機で読み取らせるための文書認識システムの研究が盛んに行なわれている．このようなシステムを実現するためには，前処理として文書の傾き検出・補正，文書を文字，図表，写真領域等に分割する領域分割，文章中の文字切り出しと認識等の様々な処理が必要となる．従来，ＯＣＲ（光学的文字読み取り装置）の研究は古くから盛んに行なわれてきたが，この技術は個別の文字を認識することに主眼を置いており，文書１ページが与えられた時に，それを構成要素に応じて自動的に機械可読形式に変換することはできない．またこれまで提案された多くのＯＣＲアルゴリズムも，実用面から見ると，様々なフォント，サイズの文字が混在し，しかもあまり印字品質の良くない日本語印刷文書に対しては十分な性能を持っているとは言い難い．

本稿では，これまで独自に開発を進めてきた文書の傾き検出・補正，領域分割，文字切り出しアムゴリズムを結合し，さらに既存のＯＣＲ手法の幾つかについて，文字認識率と読

* キャノン株式会社
** 情報工学科 助教授

み取り速度の両方をできるだけ実用的に保つための最適な組み合わせについて検討し，これらを総合して日本語印刷文書読み取りシステムを試作したのでその結果を報告する．本システムはX－ウィンドウシステム上で実現され，良好なＧＵＩ（グラフィカル・ユーザ・インタフェース）を持つと共に，移植性に優れたものとなっている．

## 2. システムの構成

日本語文書読み取りシステム(以下，システムと略す)は，図１のように画像入力部，領域分割部，文字切り出し部，文字認識部の４つの処理から構成されている．以下では，文字認識部の検討について詳細に解説し，他の処理については概略を述べるのみにとどめ，詳細については既に発表済の文献に譲る[1)2)]．

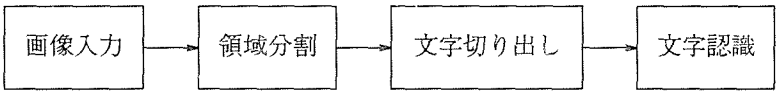



図１ システム構成

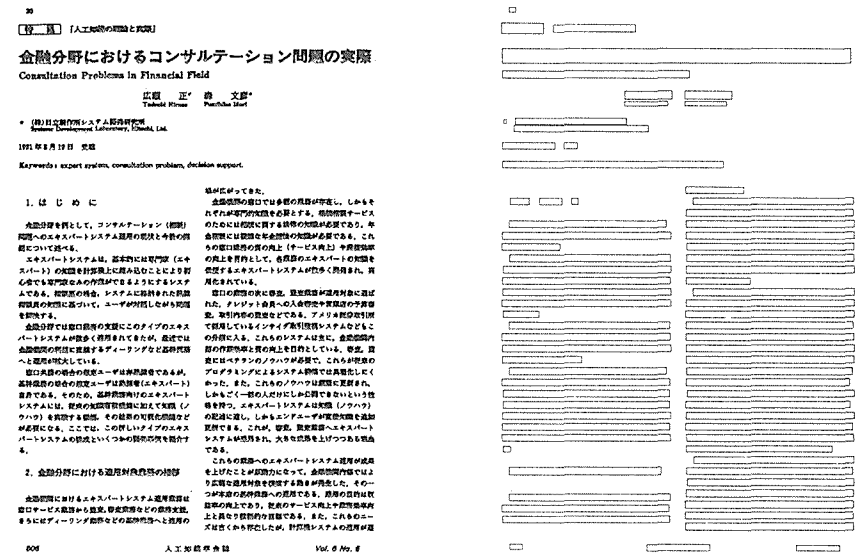
### 2.1 画像入力部

入力部では，スキャナから画像を取り込み，圧縮された形式でファイルに出力する．スキャナでは文書は 320dpi で読み込まれ，ラスタ形式の出力が得られるが，このままでは，後の処理や保存に適さないので，圧縮効率の良い SB 符号[5)] (A4 サイズの文書１ページあたり約 100KB～250KB) に変換し保存される．

### 2.2 領域分割部

領域分割部では，文書を文字，図，表，写真，フィールド・セパレータ等の構成要素からなる各領域に分割する．本システムは，主に横書きの技術文書を対象としているが，この領域分割処理は横書き／縦書きにも対応しており，また新聞等のかなり複雑なレイアウト構造を持つ文書も取り扱うことができる．領域分割結果の例を図２に示す．

スキャナから入力された文書がかなり傾いていた場合は，領域分割が正しく行なわれないため，傾きの角度の検出と補正を行なう必要がある．本処理では，領域分割前に常に傾き検出・補正を行なうことも考えられるが，補正にはかなりの時間を要するため，現システムでは領域分割結果を人間が確認し，補正を行なうか否か判断している．



金融分野におけるコンサルテーション問題の実際
Consultation Problems in Financial Field

1. はじめに

2. 金融分野における適用対象特性の抽出

図 2　領域分割の例

## 2.3　文字切り出し部

文字切り出し部では，領域分割によって得られた文字領域の各行から，個々の文字を抽出する．読み取りの対象となる文字領域およびその順序は，文書の論理構造認識を行ない自動的に指定されることが望ましいが，これは文書読み取り後のアプリケーションにも依存し，困難な課題を多く含んでいるため現システムでは実現されていない．このため，認識したい文字領域はディスプレイ上のカーソルを用いて，対話的に入力するシステムとなっている．

日本語文書では，分離文字の存在やいわゆる半角文字と呼ばれる英数字の存在が文字切り出しを困難にしている．本手法では，この問題に対処するため文献 (6) の手法を拡張し，文字サイズおよびピッチの推定を行ない，行頭と行末から再帰的に信頼度の高い場所から個々の文字を切り出す方法を用いている (図 3 参照)．ここではまず最初に，文字特徴の明確な句読点，ピリオド，ハイフン，括弧等が抽出される．次に推定サイズに収まる文字の中，分離文字でないものと比較的特徴の明確な「い」や「か」が抽出される．この処理を繰り返すことにより，大抵の全角文字は正しく抽出されるが，半角の英数字は文字サイズの推定矩形に正しく収まらないため検出が可能となる．

文字サイズの推定は，文書中で様々なサイズの文字が使われる可能性があるため，領域分割で得られた行矩形の情報を基に，領域ごとに推定される．文字の横幅の推定には行の

高さの最頻値を用いている．最大値としなかった理由は，接触などによる領域分割の失敗を考慮しているからである．また，日本語文字が2文字続くと推定された場合には，それらの文字を囲む矩形の中心を結ぶ距離を用いてサイズおよびピッチの再計算を行ない，より正確な値が得られるように考慮している．


図 3　文字の切り出し

## 2.4　認識部

文字認識部では，あらかじめ作成された標準パターンの特徴辞書と入力パターンの特徴量とのマッチングが行なわれ文字認識が行なわれる．漢字の認識では文字種が多いため，全ての標準パターンとマッチングを行なうと時間がかかる．このため図4に示すように，漢字パターンの大まかな特徴を用いて粗く分類し，次いで細かい特徴によって識別する2段階の認識手法がよく用いられる．本システムでも粗分類用として候補文字選択用辞書，詳細分類用として個別文字認識用辞書，および認識補助用辞書を作成している．

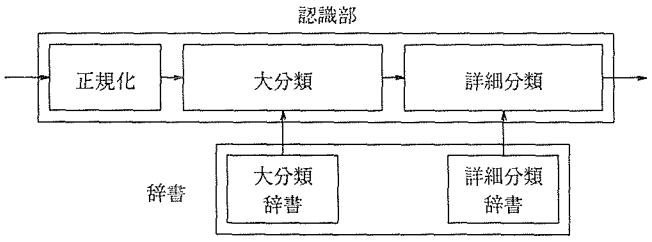
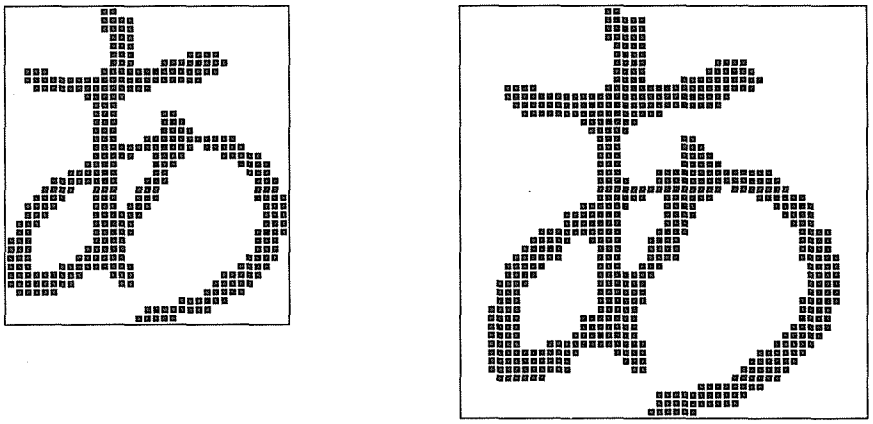



図 4　認識部の処理

### 2.4.1 正規化処理

切り出された文字画像は，標準パターンの文字と比べて、大きさや文字を囲む矩形中での位置が異なるため，統一した特徴量を抽出することができない．正規化処理では，大きさと位置の正規化を行なっている．

大きさの正規化では，白黒画素の比率が原画像に近く，斜めの線が比較的滑らかになる距離反比例法[7]を用いて，縦横比を変えずに正規化サイズ $(48 \times 48)$ になるように拡大・縮小している．ただし，拡大・縮小によって著しく形状が変形してしまうような文字や記号については，正規化を行なわず大きさの情報を考慮して認識している．また，位置の正規化では，通常の文字は上下・左右の幅を計算して正規化サイズの枠の中心に位置するように配置される．正規化の例を図5に示す．



原画像　　正規化後の画像

図5　正規化の例

「・」や「.」などのように特殊な記号については，文字行中の位置情報が重要となる．このような文字や記号については，図6に示すように，行矩形を基準とした上・中・下の位置情報を用いて認識が行なわれる．

図6　行のイメージ

### 2.4.2 大分類

JIS 第一水準に規定されているものだけでも 3,000 種類以上の文字が存在する日本語文字認識において，全ての文字を対象に詳細分類をすることは，処理の効率化の面で合理的ではない．詳細分類よりも簡単な方法で認識の対象となる文字を少なくすることができれば，効率的に文字認識ができる．また，形状の簡単な文字や記号については，特徴抽出を行なわなくても文字や記号を特定できる．

大分類処理では，正規化によって得られた文字画像から特徴抽出を行ない，あらかじめクラスタリングによって作成したクラスタの代表と特徴空間上で類似度を計算し，それが最大となるクラスタを選択している（図 7 参照）．ここで使用する特徴量は，文字パターンの細かな変動に影響されず，かつ高速に計算できるものが望ましい．このため，これまでに提案されたＯＣＲアルゴリズムの中から以下に述べるメッシュ特徴，ペリフェラル特徴[8]を併用することとした．

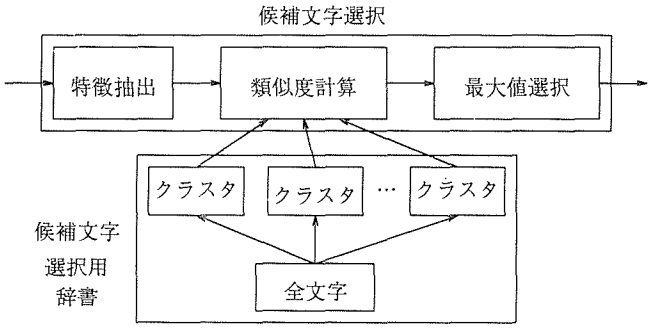



図 7　クラスタリングによる大分類

クラスタリングには，実験の初期には K 平均アルゴリズムを使用していたが，クラスタごとの偏りが大きく効率的ではなかったため，その改良版である Isodata アルゴリズムを用いている．

#### (1) メッシュ特徴

メッシュ特徴とは，正規化後の画像を 8 × 8 のメッシュで分割し，分割した各メッシュの黒画素の割合をベクトルで表した特徴量である．ベクトルの次元数は 64 次元と低次元ではあるが，黒画素の分布の概要が容易に抽出できる利点がある．メッシュ特徴の概念を図 8 にしめす．

#### (2) ペリフェラル特徴

ペリフェラル特徴とは，正規化後の画像を周辺部から走査したとき，$n$ 番めの黒画素にぶつかるまでの長さを区間ごとに積分しものを要素とするベクトル量である．本システム

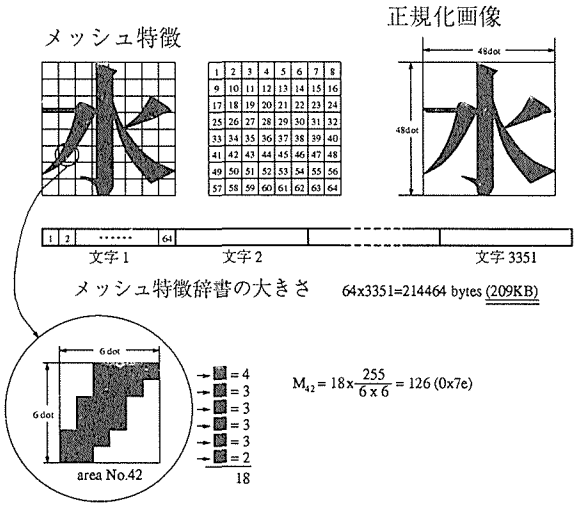



図 8　メッシュ特徴

では $n = 2$ まで求め，1 辺を 8 分割するため，ベクトルの次元数は 64 次元となる．ペリフェラル特徴の概念を図 9 にしめす．

### (3) 類似度

類似度とは，2 つのベクトルの類似性を示す量であり，2 つのベクトルの内積をそれぞれのノルムにより規格化した値である (式 1)．ベクトル間の角度が類似性を示しており，長さによる差は現われない．

$$d_r = \frac{\boldsymbol{X} \cdot \boldsymbol{Y}}{||\boldsymbol{X}|| \cdot ||\boldsymbol{Y}||} \tag{1}$$

### 2.4.3 詳細分類

詳細分類では，大分類で選ばれた文字の細かな違いを検出し，正確に識別する必要がある．そのためには，高次元の特徴量を用いて候補文字との類似度を計算し，最も信頼度の高い文字を選択する必要がある．ここでは，過去の文献，資料から比較的認識率の高いと思われる以下の 2 通りの手法を選択し，読み取り実験を行なって本システムに最適なものを採用した．

### (1) PDC 特徴

外郭方向寄与度 (PDC) 特徴は，萩田等[3]が 1983 年に提案した特徴量で，図 10 に示すように，ある点から連続する黒画素の量を調べ，全方向に対する各方向の黒画素連結量の割

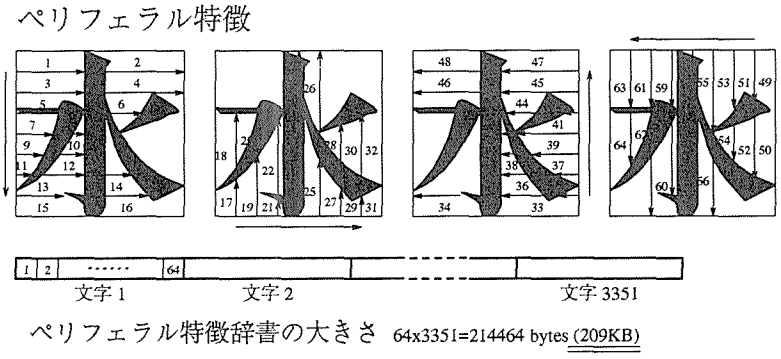



図 9 ペリフェラル特徴

合を方向寄与度特徴とするとき，周辺部から走査して出会った点 (外郭点) における方向寄与度を要素とするベクトル量である．

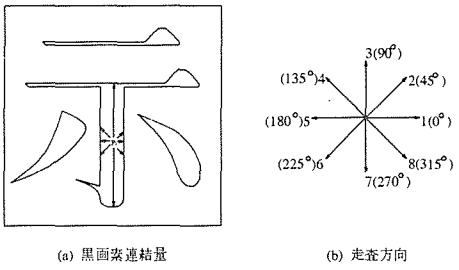



図 10 PDC 特徴

### (2) 方向線素特徴

方向線素特徴は，孫等[4]が提案した特徴量で，図 11 に示すように原画像を細線化し，各画素に方向を与えたものを方向線素とするとき，各メッシュ中の方向線素の数を方向ごとに取り出した特徴量である．メッシュの境界部に存在する線素は多少の位置ずれに対しても大きく変化してしまうため，メッシュは重なるようにとり，その分，メッシュの中心部に重みを与えている．

## 3. 実 験

まず個別文字認識手法を決定するために予備実験として，前に述べた PDC 特徴を使用する方法と方向線素特徴を使用する方法を比較・検討した．ここでは，両手法の単独の認識

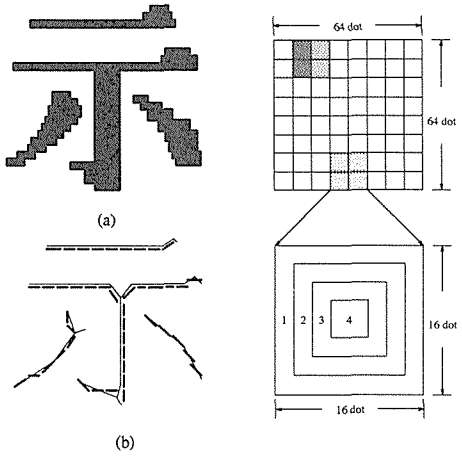



図 11　方向線素特徴

性能，処理速度を比較するため，大分類を行なわずに認識実験を行なった．この実験では，各手法とも最適な認識率が得られるよう，辞書の作成，認識パラメータの調整を行なったシステムを作成し，性能を評価している．

### 3.1　実験環境

実験環境を図 12 に示す．実験には，Sun のワークステーション SPARCstation2 を使用し，画像はイメージスキャナ EPSON GT-8000 から 320dpi の解像度で入力した．実験に使用した文書は学会論文誌や一般の雑誌で，約 48,000 文字を認識させた．文字読み取り実験の様子を図 13,14 に，実験結果の一部を図 15 に示す．なお，今回の実験で対象とした文字は，ロシア文字を除く JIS 第 1 水準文字 3,351 文字である．

### 3.2　実験結果および検討

表 1　PDC 特徴と方向線素特徴の比較

| 使用特徴量 | 認識速度 [文字/秒] | 認識率 |
|---|---|---|
| PDC 特徴 | 0.47 | 96.14 |
| 方向線素特徴 | 0.99 | 90.06 |

表 1 から明らかなように，PDC 特徴と方向線素特徴の性能比較では，PDC 特徴の方が認識率が高いが認識速度は遅いことが確認された．この結果，大分類と組み合わせた場合

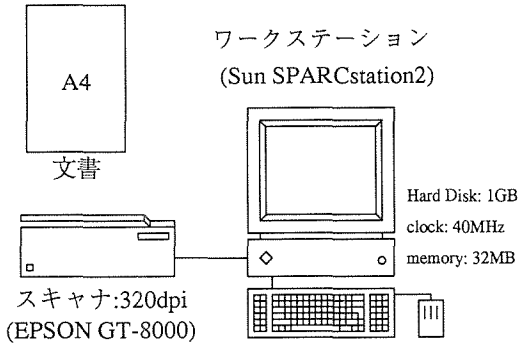



図 12 ハードウェア構成

でもかなりの認識率（約 95%以上）を確保するため，詳細分類としては PDC 特徴を採用することとした．

以上の予備実験から，大分類としてメッシュ，ペリフェラル特徴の併用，詳細分類として PDC 特徴を用いたシステムを再構築し，各種パラメータの最適調整を行なった後，様々な文書に対して読み取り実験を行ない，システム全体の評価を行なった．各文書に対する文字認識率（文字切り出しの誤りを除く）は表 2 に示す通りである．以下に本システムの性能評価について述べる．

表 2 文書の種類による認識率の変化

| | | |
|---|---|---|
| 1 | 信学会論文誌，人工知能，bit，日経エレクトロニクス | 99% ～ |
| 2 | LaTeX テスト用文書 (12 ポイント以上) | 約 99% |
| 3 | 電子情報通信学会誌 | 約 99% |
| 4 | インターフェイス | 約 98% |
| 5 | 情報処理学会論文誌，情報処理 | 約 97% |
| 6 | 信学会技術研究報告，情報処理学会研究報告 | 94% ～98% |
| 7 | LaTeX テスト用文書 (10 ポイント以下) | 92% ～97% |
| 8 | UNIX MAGAZINE | 93% ～95% |
| 9 | PIXEL | 92% ～94% |

表 2 に示す通り，電子情報通信学会論文誌と電子情報通信学会誌ではいずれも 98.5% 以上の認識率が得られた．両者とも使用されているフォントはよく類似しており，辞書の作成に用いたフォントと類似していたのが原因と思われる．人工知能学会誌，bit，日経エレクトロニクスでは，使用されているフォントがいずれも電子情報通信学会論文誌のものと類似しており，かなり安定した高い認識結果（99%以上）が得られた．

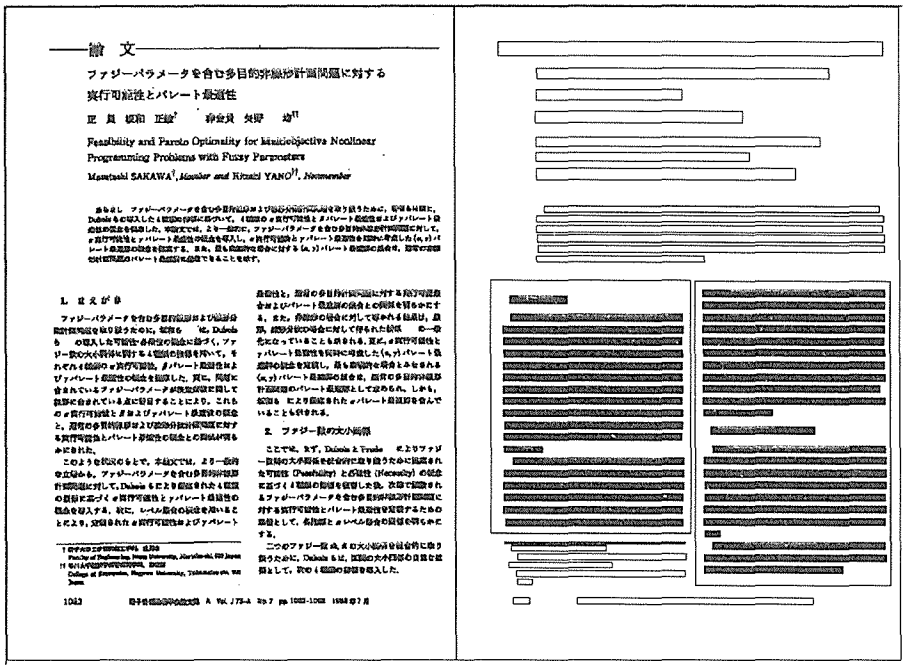


図 13　読み取り実行中の画面 (認識する領域の指定)

また，情報処理学会論文誌と情報処理（会誌）では全く同じフォントが使用されているが，電子情報通信学会論文誌などに使用されているフォントとは明らかに異なり，「こ」，「に」，「た」に共通する，上下に位置する横線は全て連結していて丸みも少ない．そのため，やや認識率が低く，97% 前後になっている．

電子情報通信学会技術研究報告と情報処理学会研究報告は，投稿者が独自のワープロなどで出力した原稿をそのまま載せているため，かなり低品質なものになっている．このような理由で認識率も 94% から 98% と幅広く，平均的に見ると 96% 程度である．

UNIX MAGAZINE は縦長で文字線の長さが比較的短いスマートなフォントを使用しており，このシステムとはあまり相性が良くなく，93% から 95% という結果になっている．インターフェイスは，UNIX MAGAZINE とほとんど同じフォントを使用しているが，印刷の質が良いためか，98% の認識率を示した．

PIXEL は電子情報通信学会論文誌と類似のフォントであるにも関わらず，悪い結果を示している．この原因は文字の大きさが小さく，横線の多くが「かすれ」を起こしていたからだと思われる．

LaTeX で作成したテスト用文書では，10 ポイント以下の文字とゴシック体で認識率が低かったが，12 ポイント以上の明朝体に関しては安定して 99% を達成している．300dpi の

## 1. まえがき

ファジーパラメータを含む多目的線形および線形分数計画問題を取り扱うために，坂和ら　　は，Duboisら　　の導入した可能性・必然性の概念に基づく，ファジー数の大小関係に関する4種類の指標を用いて，それぞれ4種類のα実行可能性，βパレート最適性およびγパレート最適性の概念を提案した．更に，問題に含まれているファジーパラメータが決定変数に関して線形に含まれている点に着目することにより，これらのα実行可能性とβおよびγパレート最適性の概念と，通常の多目的線形および線形分数計画問題に対する実行可能性とパレート最適性の概念との関係が明ら

図 14　原文書の例 (切り出しの結果も表示)

レーザープリンタによる出力なので印字品質が悪く，10 ポイント以下では画数の大きい文字がほとんど「つぶれ」ていた．これは TEX のフォントの質によるものである．

以上表 2 から明らかなように，1 位から 5 位の文書に関しては，ほとんど問題なく実用的に使用できる．6 位から 9 位の文書は，認識率にバラツキが有るが，文字の大きさや印字品質などに注意して使用すれば実用的な価値はかなり高いものと思われる．

また，システムの別な評価尺度として認識速度があげられるが，認識速度については 1 秒当たり約 3 文字で安定していた．文書の種類による認識速度の差は観測されなかった．本システムはプログラムの最適化を行なっていないため，多少認識速度を改善できる余地は残っているが，全てをソフトウェアのみで構成しているため，大幅な認識速度の向上は望めない．しかしながら，最近のパーソナル・コンピュータ，ワークステーションのＣＰＵ速度の向上は目覚しく，移植性を重視した本システムの有効性も高まるものと思われる．

## 4. むすび

これまで多くのＯＣＲアルゴリズムが提案されてきたが，文書読み取り装置として構成し，現実の様々な印刷文書に対して読み取り実験を行なった例は少ない．本研究は，イメージスキャナーをワークステーションに接続し，文書のスキャンから領域分割，文字領域の読み取りまでの一連の処理を行なう日本語文書読み取りシステムの試作を行なって，その

□まえ□□き
ファジーパラメータを含む多目的線形および線形分数計画問題を取り扱うために，坂和らは，Ｄｕｂｏｉｓらの導入した可能性・必然性の概念に基づく，ファジー数の大小関係に関する４種類の指標を用いて，それぞれ４種類の $\alpha$ 実行可能性，$\beta$ パレート最適性および γ パレート最適性の概念を提案した．更に，問題に含まれているファジーパラメータが決定変数に関して線形に含まれている点に着目することにより，これらの $\alpha$ 実行可能性と $\beta$ および γ パレート最適性の概念と，通常の多目的線形および線形分数計画問題に対する実行可能性とパレート最適性の概念との関係が明ら

図 15　認識結果の例 (誤認識文字は反転表示)

性能を評価したものである．このようなシステムを実現するためには，文書画像処理分野の様々な研究・開発が要求されるが，本システムではこれまで当研究室で開発してきた関連分野の手法を総合し，全てソフトウェアで実現している．本稿では，特に文字認識部の構成について詳細に検討しているが，大分類としてメッシュ，ペリフェラル特徴を用い、詳細分類では PDC 特徴を用いた文字認識が有効であり，また大分類辞書を構成する際のクラスタリング手法としては，Isodata アルゴリズムによるクラスタリングが有効であることが示された．その結果，おおよそ認識率 97%, 1 秒当たり 3 文字という認識速度を実現するシステムをワークステーション上に構築することができた．また，学会誌や一部の雑誌に対しては，99% 以上の認識率が得られることが確認された．

今後の課題は，低品質の印刷文字に対して良い結果が得られるような認識手法の確立と認識速度の向上，および文書の論理構造理解等の手法の確立である．

## 参 考 文 献

1) M. Okamoto and A. Miyazawa: “An Experimental Implementation of Document Recognition System for Papers Containing Mathematical Expressions”, Structured Document Image Analysis, Springer-Verlag, pp.36-53(1992).

2) M.Okamoto and M. Takahashi: “A Hybrid Page Segmentation Method”, Proceedings of Second International Conference on Document Analysis and Recognition, IEEE Computer Society, pp.743-748(1993).

3) 萩田, 内藤, 増田: “外郭方向寄与度特徴による手書き漢字の識別”, 電子通信学会論文誌, J66-D, No.10(1983).

4) 孫, 田原, 阿曽, 木村: “方向線素特徴量を用いた高精度文字認識”, 電子通信学会論文誌 (D-II), Vol.J74-D-II, No.3, pp.330-339(1991-3).

5) 岡本, 高橋, 河田: “セグメント・ブロック符号を用いた画像の拡大／縮小, 回転, 境界追跡 アルゴリズム”, 電子通信学会論文誌, Vol.J69-D, No.7, pp.1075-1082(1986).

6) 宮原, 木村, 豊田, 宮田: “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”, 電子通信学会論文誌, J72-D, No.6, pp.846-854(1989-6).

7) 正嶋, 葛貫, 中島, 坂東, 平沢: “二値画像の各種拡大／縮小方式の性能評価及び処理速度改良方式”, 情報処理,Vol.26.No.5,pp.920-925(1985-9).

8) 梅田: “マルチフォント印刷漢字の分類”, 電子通信学会論文誌, J62-D,No.2, pp.133-140(1979-2).