日本語活字文書画像を対象とする文字列検索手法のフォントロバスト性の検証

新國　三千代・田中　譲

This paper evaluates the font-robustness of a full-text search method for Japanese document images on the "Transmedia System". This method encodes each Japanese character image to a 'pseudo code' using the "Peripheral Features". It is capable of searching document images for an arbitrarily given character string. In the previous paper, the authors showed that it has very high precision in the experiment using 200dpi Japanese single-font document images of 10.5 point MS-Minchou font. In this paper, the authors verified that it achieves the 99.7% average recall ratio, the 85.2% average precision ratio and 0.148 noise factor for about 1,600 two-Kanji-character terms. This experiment used 400dpi multi-font document images, each including about 10,000 Japanese characters of 10.5 point with equal frequency of MS-Minchou and MS-Gothic.

1．はじめに

筆者等は，前論文（新國・田中，2001）において，日本語活字文書画像中の文字列を検索する方法を提案した．これまでの日本語の文書画像中の文字を検索する研究には，OCRの文字認識を利用したものがある．例えば，文書画像をテキストコードに変換する際の認識誤りの特性を用いた全文検索（丸川他，1995）や，確率的な全文検索方法（太田他，1998）などである．これらの研究の主たるテーマは検索文字の再現率を向上させることである．前者では99.5%，後者では99.26%に改善されることが言及されている．

一方，筆者等はトランスメディアシステムと呼ばれる，OCRの文字認識とは全く異なる方法で文書画像中の文字を検索する試みを行っている（田中知朗他，1997；遊佐・田中，1994；遊佐・田中，1995）．これは，文書画像中の個々の文字の形状特性から特徴量を求めて擬似コード化し，この擬似コードを用いて文字列検索を行うというものである．この方法では，検索時の再現率をほぼ100%にすることが可能である．また，疑似コード化は文書画像に対して機械的に行われるため，スキャナー入力以外に人手をかける必要もない．更に，辞書等を用いて文字認識を行う訳ではないので，辞書が存在しない言語に対しても適用できるという利点がある．

筆者等は，前論文で，日本語活字文書の文字の特徴量として文字の周辺情報に基づくペリフェラル特徴（梅田，1979）を導入し，これに基づき各文字の擬似コードを生成して，文書画像中の文字列を検索する方法を提案した．そして，この疑似コードを用いて文字画

Nikkuni Michiyo　札幌学院大学社会情報学部
Tanaka Yuzuru　北海道大学知識メディアラボラトリー

像同士の距離を定義し，距離的に近い文字画像を同じ文字と見なすことにより文字列検索を行っている．ペリフェラル特徴を採用した理由は，これがマルチフォントロバストな特徴として提案されたものであるため，標準体以外にゴシック体などが混在する通常の文書画像を扱う上で最適と考えたためである．

前論文では，10.5ポイントのMS明朝の単一フォントで書かれた日本語活字文書を200dpiという低解像度でスキャナー入力した文書画像に本手法を適用した結果，2文字漢字検索で平均再現率100%時に高い適合率が得られることを示した．本稿では，日本語活字フォントの中で最も形状が大きく異なる標準体とゴシック体が混在するマルチフォント文書画像を対象に，筆者等が提案する検索手法のフォントロバスト性を検証する．通常の文書では，標準体にゴシック体が混在する場合が多いことから，形状の相違が大きいこの2つのフォントが混在する文書についてフォントロバスト性が検証できれば一般的な文書画像の検索にも十分対応できると考えられる．一般に，文書中に含まれるゴシック体はタイトルや強調文字などで使用されるため，多くても1～2割程度である．しかし，ここでは1行おきにMS明朝とMSゴシックを混在させて，少なくとも同じ文字が標準体とゴシック体の両方を1個以上有するという厳しい条件下で実験を行うことにした．実験では標準体としてはMS明朝，ゴシック体にはMSゴシックを使用した．

このようにして作成されたマルチフォントの日本語活字文書に対して筆者等の検索手法を適用した結果，10.5ポイントの文字の大きさで解像度が400dpiのとき2文字漢字検索で十分実用に耐え得る検索精度が得られることが分かった．評価実験では，前論文と同じ科学分野の抄録集1万字程度の文書画像を対象に約1,600組の2文字漢字検索を行い，平均再現率，平均適合率，そして非適合係数（＝Σ非適合件数/Σヒット件数）を求めて評価を行った．

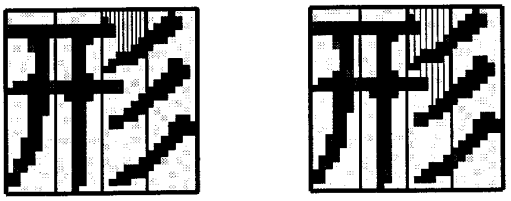
以下，2．で前論文で提案したペリフェラル特徴を用いた文字画像の擬似コードの生成方法と文字画像の検索手法について簡単に紹介し，3．で本稿で使用する検索精度の評価尺度について述べ，4．で評価実験の準備と方法，5．で評価実験の結果と考察，6．で結論を述べる．

2．ペリフェラル特徴を用いた文字画像の擬似コードの生成と文字画像の検索手法

ペリフェラル特徴を用いた文字画像の擬似コードの生成と文字画像の検索手法については前論文（新國・田中，2001）で詳しく述べているので，本節では簡単に紹介することにする．

ペリフェラル特徴とは，“文字の外郭形状の特徴を面積に置き換えたもので，文字の外接方形領域を切り出し，四つの外接枠をそれぞれn分割し，分割された外接枠の各部分から反対側の外接枠方向に文字部に出会うまでの文字部でない領域の面積を計数して，これを当該分割領域の面積で割って規格化する”（梅田，1979）ことにより求められる．その結果得られるペリフェラル特徴は0～1の値になる．図2.1のa）は，辺を4分割して上辺からペリフェラル特徴を求める例を示している．4分割では上辺，右辺，下辺，左辺と4辺×4分割＝16のペリフェラル特徴が抽出される．これを一次ペリフェラル特徴と呼ぶ．更に，口と国のように漢字の内部構造の違いを反映させるために，図2.1のb）のように最初の文字部の出合いではそのまま計数を続け，次の文字部に出会うまでの面積を計数して面積で規格化したものを二次ペリフェラル特徴と呼ぶ．

文書画像からの文字領域の切り出しは，下記の順で頁単位に行う．



a）一次ペリフェラル特徴　　b）二次ペリフェラル特徴

図 2.1 ペリフェラル特徴

1）文書領域の原点を決める．文書領域の左端の座標を原点（0，0）とする．
2）行を切り出す．
3）文字領域の左右外郭辺（文字幅に相当）を切り出す．
4）文字領域の上下外郭辺（文字高に相当）を切り出す．

1）～4）の結果，各文字領域の大きさは文字により異なってくる．また，「一」のように極端に文字の高さが低いものについては，書き出しや終端のノイズや歪みによって高さの影響が大きく出るため，便宜上文字の高さを平均の高さで補正している．3）の処理では「化」や「川」のように文字の途中で垂直方向に空白がある場合は機械的に分離されて複数文字となる．この場合は，文書画像を90度回転して文字の高さを求め，これを文字幅にすることで本来の文字幅を求めることができる．しかし，本研究の実験に用いた文書では分離文字は52種で全体の5％と少ないことと同じ文字が同様に分離されていれば検索上は問題ないことから，特にこのような操作は行わずに分離文字は分離した形でそれぞれの文字領域を切り出している．

文字領域の分割では，上下左右の各辺をn分割する際に，1）割り切れない，2）分割した数がn個に満たない，3）全ての分割領域の大きさが均等にならない，といったことが生じる．この場合は，1）小数点以下切り捨てた値を分割幅とする，2）満たない分の値を0とする，3）n番目の分割幅は残った幅とする．この原則をすべての文字に対して適用すれば，特に問題が起きることはない．

次に，各文字画像の擬似コードを生成するために，まず各分割領域毎にペリフェラル特徴を求め，ペリフェラル特徴の値0～1を0～255対応させる．そして，横軸が0～255，縦軸は横軸の値をもつ文字の数からなるヒストグラムを作成する．例えば，文字の各辺を4分割すると，一次ペリフェラル特徴では16個のヒストグラム，二次ペリフェラル特徴も含めると32個のヒストグラムが作成される．当然ヒストグラムの形は辺の分割領域によって異なるが，図2.2のように各ペリフェラル特徴のヒストグラムにおいて面積がほぼ同じになるようにn^2等分して，ヒストグラムの各部分領域に含まれる各文字にnビットからなる2^n個のコード（n＝2の場合は，00～10の2ビットコード，n＝3ビットの場合は，000～111の3ビットコード）を付与する．この各コードを単位コードと呼ぶ．このとき，ミラーコーディングにより互いに隣り合う領域に属する文字は1ビットの違いしか持たないようにする．以上の結果，ヒストグラムの各部分領域に含まれる文字には全て同じコードが付与される．このようにして求めた単位コードを特徴量の数だけ横に並べて1つの文字に対応する擬似コードを生成する．この擬似コードを用いて検索したい文字画像と照合される文字画像との間に距離を定義し，距離的に近い文字，すなわち検索文字との距離が，ある許容距離内にある文字を同じ文字とみなすことにより文字の検索を行う．検索文字に指定された1文字画像

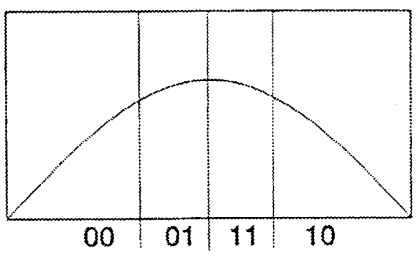


図 2.2 ミラーコーディング

Aと検索対象の文書画像中の照合される1文字画像Bとの間の距離の定義は次の通りである．但し，各特徴量に連続番号を付与しているものとする．

AとBの第i番目の特徴量に対応する単位コードをa_i, b_iで表し，

$$c_i \begin{cases} \equiv 0 : a_i \text{と} b_i \text{が} i \text{番目の特徴量のヒストグラムで同じ部分領域に含まれる} \\ \equiv k : a_i \text{と} b_i \text{が} i \text{番目の特徴量のヒストグラムで} k \text{個離れた部分領域に含まれる（隣り合う場合を1とする，} k = 1 \sim 2^n - 1 \text{、} n \text{は単位コードのビット数）} \end{cases}$$

とするとき，

AとBの距離d_{AB}を$d_{AB} = \sum_{i=1}^{m} c_i$（$m$＝ペリフェラル特徴の数）と定義する．

a_i, b_iが近い部分領域に含まれていればd_{AB}は小さな値に，離れた領域に含まれるほど大きな値になる．スキャナで入力した文書画像にはノイズや歪みが入るため，AとBが同じ文字であってもその擬似コードが同じになることは極めて希である．従ってこの値が0になることは殆どない．

検索時には，指定された検索文字の文字画像Aと照合対象となる全ての文字画像Bとの距離d_{AB}を求め，この距離がある値（許容距離と呼ぶ）以下であるものを同じ文字と見なす．2文字以上を検索時に指定した場合は，各文字についてそれぞれの距離がある許容距離以下のとき，これを同じ文字列と見なす．検索文字数が増えると1文字の場合に比べ，検索文字で指定した文字列に近い距離を持つ全く異なる文字列が見つかる確率は当然小さくなることから検索の精度は上がることになる．

検索時の検索文字の指定は，文書画像中の文字画像すなわち擬似コードを直接指定して行うか，キーボードから入力した文字列を画像化して検索対象の文書画像で作成したヒストグラムを用いて擬似コード化したものを指定することにより行う．

3．検索精度の評価尺度

各検索文字に対するヒット件数と適合件数および非適合件数を次の通り定義する．

ヒット件数＝同じ文字とみなされて検索された件数（非適合件数含む）

適合件数＝正しくヒットした件数

非適合件数＝ヒット件数－適合件数

検索精度の評価は，次の3つの評価尺度を用いて行う．

平均再現率＝（Σ適合件数/Σ検索されるべき対象文字の出現数）×100

平均適合率＝（Σ適合件数/Σヒット件数）×100

非適合係数＝Σ非適合件数/Σヒット件数

文書画像中に2個以上出現するある長さをもった文字列n組について，各文字組のヒット件数，適合件数，非適合件数を求めてそれぞれの総和を求めてから除する．これは次の理由による．通常ある目的を持って書かれている日本語の文書では出現頻度の高い文字列は少ない．例えば，実験で用いる約1万字の文書では，隣接する漢字2文字の出現頻度は18以下が全体の81%を占める．10以下は，全体の約62%を占めるが，この中では頻度1が約16%で最も多く，次は頻度2の約14%，頻度3～7は合わせて約26%である．仮に検索したい文字列の出現頻度を2とすると，別の文字列が1個以上ヒットしてくると，適合率は66.67%以下になる．しかし，頻度が10の場合は適合率は90.91%以下となり，非適合件数1件の持つ重みが出現頻度により大きく異なってくる．再現率についても同様のことが言える．すなわち，出現頻度が小さい場合は再現率や適合率に与える影響が大きくなる．このような出現頻度の影響を避けるためにそれぞれの総和を求めてから割っている．前論文では，各検索文字毎に再現率と適合率を求め，最後にそれらを足して

平均再現率と平均適合率を求めたが，今回の実験では出現頻度の違いを排除するためにこのような方法をとった．因みに，今回の方法を前回の実験に適用すると，平均再現率，平均適合率とも若干ではあるが高くなる．

非適合係数とは，検索時にヒットした件数の中で適合しない文字列がどの程度ヒットしてくるかを示す値である．仮にヒット件数が10のときにこの値が0.1であるとすると，1個の違った文字列がヒットしていることを意味するが，この値が小さければ実用上問題はないと考えることができる．一般にスキャナ入力時のノイズの影響は頻度が多いほど大きくなると予想されるが，検索時にはこれらがある範囲内に収まっていれば同じ文字とみなされるので，特異なノイズが出現しない限り，出現頻度に反比例して平均適合率が低くなるというわけではない．逆に，出現頻度が小さくてもよく似た文字がたくさん出てくれば平均適合率は低くなる．実験の評価では，平均再現率が99.5%以上でかつ再現率が100%未満のものが数%以内，そして非適合係数が0.2未満であれば実用上問題のない範囲であると考えることにする．

4．評価実験の準備

ここでは，実験で使用する文書画像と疑似コードの生成，そして実験の方法について述べる．

4.1 使用する文書画像

実験に使用する文書画像の文字数と擬似コードの関係については既に前論文で検証済である．すなわち，英数字が混在する科学技術分野の日本語抄録集の場合，総文字数が5千文字以上あれば擬似コードを生成する上で問題がないことが示されている．前論文の実験では，天文学会のホームページに公開されている年会講演の抄録集から総文字数約4万文字の抄録を抽出して擬似コードを生成し，その内の約1万文字を対象に検索精度を検証している．本稿においても，前論文で使用した同じ文書を実験に用いることにする．

さて，本論の課題は筆者等が提案する文字列検索手法のフォントロバスト性を検証することである．すなわち，複数の異なったフォントが混在する文書画像を対象に検索を行ったとき，異なるフォントの文字も検索可能なことを示すことである．ところで，ペリフェラル特徴はフォントロバストな特徴量として考案されたものなので，当然異なる文字フォントが混在する文書画像にも適用可能である．特に，活字フォントの中で最も形状が大きく異なる標準体とゴシック体が混在するようなマルチフォント文書画像に適用した場合にもそのフォントロバスト性を保証する必要がある．実際の文書では，標準体にごく僅かのゴシック体が混在する場合が殆どなので，形状の相違が大きいこの2つのフォントについて検証できれば一般的な文書画像の検索にも十分対応できると考えられる．一般に，文書中に含まれるゴシック体は多くても1～2割程度であるが，ここでは1行おきにMS明朝とMSゴシックを混在させて，少なくとも同じ文字が標準体とゴシック体の両方を1個以上有するという厳しい条件を作って実験を行うことにした．実験に使用する約4万文字の抄録文書を1行おきにMS明朝とMSゴシックにして，機械的に隣接する2文字を抽出し，頻度2以上になる2文字漢字を調べると，MS明朝とゴシックの両方のフォントを有するものは全体の約85%になる．従って，この方法により厳しい条件は維持できていると考えてよい．

前論文では10.5ポイントの単一フォントの文書を対象にして200dpiという低解像度（1文字平均の幅と高さの画素数は25×22）でスキャナ入力した文書画像を実験データとして用いた．前回の実験で解像度を上げると検索精度が上がることを確かめていたので，本稿では異なったフォントを対象にしている

ことから解像度（画素数）を上げて実験を行った．表4.1に示すように，文字サイズを10.5ポイントと12ポイントにしたマルチフォント文書を200，300，400dpiでスキャナ入力して文書画像を作成し，外郭辺がほぼ四角形になるMS明朝の〝形〟という文字で画素数（以下，幅×高さで記す）を比較すると，10.5ポイントの200dpiは，26×24画素，300dpiでは39×37画素，400dpiでは53×50画素となる．12ポイントの200dpiは，30×27画素，300dpiは45×42画素，400dpiは61×57画素である．

以上まとめると，実験で使用する文書は，天文学会のホームページに公開されている年会講演の抄録集から抽出した抄録で，出現する文字種は，ひらがな，カタカナ，漢字，英数記号合わせて1,035字（異なる文字種の数），総文字数は38,566字である．この文書を1行おきにMS明朝とMSゴシックを混在させて，10.5ポイントと12ポイントのマルチフォント文書を作成し，600dpiでプリンタ出力する．次に，これを200dpi，300dpi，400dpiでスキャナ入力して2値モノクロの文書画像を作成する．そして，この画像データを24ビットのjpegに変換する（実験で使用したシステムの制約による）．2値モノクロに比べ更に精度が落ちたデータになるが，黒画素の濃度が50%以下のものをノイズと見なして無視すると，2値モノクロと同様の文字パターンが得られるため，この変換が精度に影響を与えることはない．図4.1はこのようにして得られた10.5ポイントMS明朝とMSゴシックの〝形〟という文字画像をほぼ同じ大きさまで拡大したものである．同じフォントでも解像度により微妙に違っていることが分かる．

表4.1 文字サイズと解像度および画素数の関係

文字サイズ	解像度	画素数(文字幅×文字高)+
10.5ポイント	200dpi	26×24
12ポイント	200dpi	30×27
10.5ポイント	300dpi	39×37
12ポイント	300dpi	45×42
10.5ポイント	400dpi	53×50
12ポイント	400dpi	61×57

+ほぼ四角形になるMS明朝の〝形〟で代表

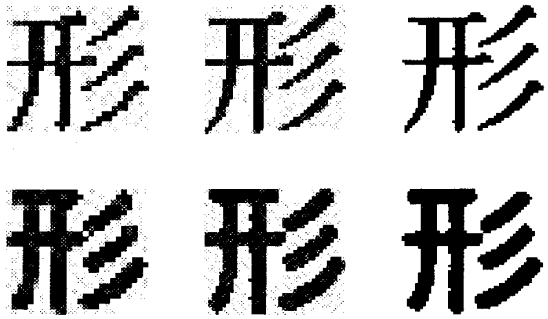


図4.1 10.5ポイント“形”の文字画像

4.2 擬似コードの生成

前論文で単一フォント文書の場合について，疑似コードを生成するために，①採用するペリフェラル特徴と②最適な単位コードのビット数および③最適な辺の分割数を調べた．その結果，①では第1と第2の二つのペリフェラル特徴を採用し，②では3ビットコード，③では200dpi（1文字の幅25×高22画素）の場合は6分割が最適であることを示した．①と②についてはマルチフォント文書の場合も単一フォント文書と同じ結果であった．しかし，③の辺の分割数については画素数（解像度）の影響を受けるため検証が必要である．表4.1で示した通り，300dpiの画素数は200dpiの文字幅と文字高のほぼ1.5倍，400dpiでは文字幅，文字高とも200dpiのほぼ2倍になっている．実験の結果，45×42画素以上の場合は，辺の分割数は8分割の方が6分割よりも検索精度が高く，これより小さい画素の場合は，6分割の方が高くなる．すなわち，10.5ポイントの文字サイズで考えると，200dpiと300dpiは6分割，

表 4.2 画素数と辺の分割数による検索精度の比較

文字サイズ	画素数+（幅×高さ）	辺の分割数	許容距離*	56/88	58/90	60/92	62/94	64/96
10.5p	39×37	6 分割	平均再現率（%）	95.83	97.10	98.14	98.84	99.36
			平均適合率（%）	90.60	85.74	79.25	71.42	62.65
	39×37	8 分割	平均再現率（%）	96.14	97.03	97.74	98.33	98.83
			平均適合率（%）	67.10	59.65	52.20	44.65	37.62
文字サイズ	画素数+（幅×高さ）	辺の分割数	許容距離*	60/78	62/80	64/82	66/84	68/86
12p	45×42	6 分割	平均再現率（%）	97.68	98.58	99.06	99.37	99.68
			平均適合率（%）	80.27	73.42	65.05	56.12	46.79
	45×42	8 分割	平均再現率（%）	97.60	98.30	98.88	99.31	99.58
			平均適合率（%）	90.83	87.15	82.99	77.92	71.86

+ほぼ四角形になる MS 明朝の〝形〟で代表 * 6 分割時の許容距離／8 分割時の許容距離

400dpi は 8 分割の方が高くなる．表 4.2 に 39×37 画素と 45×42 画素のそれぞれ 6 分割と 8 分割の検索精度の比較を示したが，単一フォント文書でもマルチフォント文書の場合と結果は同様であった．これは，ある程度の画素数がある場合は 6 分割よりも更に細かく 8 分割した方が漢字の形状の特徴をうまく抽出できることを示している．

以上のことから，擬似コードの生成は，一次と二次ペリフェラル特徴を用いて，単位コードのビット数は 3 ビット，辺の分割数は 39×37 画素以下では 6 分割，45×42 画素以上では 8 分割で行うことにする．

4.3 実験の方法

実験を行うために，最初の約 1 万文字（総文字数が 10,444 字，異なる文字数 691 字）について擬似コードと各文字画像との対応をとり，検索結果の正否を判定するプログラムを作成した．4.1 で述べた通り，実験に用いるマルチフォント文書では，頻度 2 以上の 2 文字漢字で MS 明朝と MS ゴシックの両方のフォントをもつものは 85%あった．これらの 2 つのフォントが含まれる割合は異なっているが，半々の場合とそうでない場合とで再現率や適合率に大きな相関は見られない．従って，ここでは含まれるフォントの割合について問題にする必要はない．そこで，文書中の隣り合う 2 文字漢字を機械的に抽出し，それらを検索文字列に指定して検索精度を求めることにした．

評価実験の方法は次の通りである．

1） 1 万弱の文書画像中で隣接する漢字の 2 文字画像を機械的に抽出し，分離文字でない出現頻度 2 以上の 2 文字漢字を検索用文字列とする．つまり，文書画像中の文字で検索用文字列を指定する．内部的には 2 文字の擬似コードが並んだものになる．2 文字漢字は，約 1,600 個生成される．

2） 1）で抽出した 2 文字漢字を検索文字列に指定して約 1 万文字の文書画像中の全ての隣接する 2 文字との距離を求める．

3） 各検索文字毎に許容距離を動かしながら，検索の適否を判定するプログラムを実行し，ヒット件数，適合件数，非適合件数を求める．

4） 3）の各結果の和をとり，平均再現率と平均適合率，非適合係数を求める．

実際の検索では，2 文字以上の漢字を検索文字に指定するのが一般的であるが，3 文字以上にすると，検索文字と同じ並びの文字列が出現する確率は 2 文字よりも少なくなるた

め検索精度は2文字の場合よりも高くなる．従って，実験では2文字について検索精度を評価しておけば十分である．

5．評価実験の結果と考察

本節では，4.1で述べた通りMS明朝とMSゴシックを1行おきに混在させたマルチフォント文書を対象とする評価実験の結果について述べ，更にトランスメディアの従来方式の特徴量を用いた場合と比較した結果について述べる．

5.1 MS明朝とゴシックを1行おきに混在させたマルチフォント文書画像への適用結果

図5.1は，10.5ポイントMS明朝のみからなる単一フォント文書と10.5ポイントMS明朝とMSゴシックを1行おきに混在させたマルチフォント文書を400dpiでスキャナ入力して作成した文書画像中の同じ文字の最大距離の累積分布を示したものである．マルチフォント文書画像の同じ文字の最大距離は，単一フォント文書画像の場合よりも大きくなっている．従って，平均再現率を上げるためには単一フォント文書画像の場合よりも許容距離を大きくする必要がある．これは，200dpiや300dpiの場合も同様である．

表5.1は，MS明朝とMSゴシックを1行おきに混在させたマルチフォント文書画像を対象に，文字サイズと解像度を変えて約1,600組の2文字漢字について平均再現率と

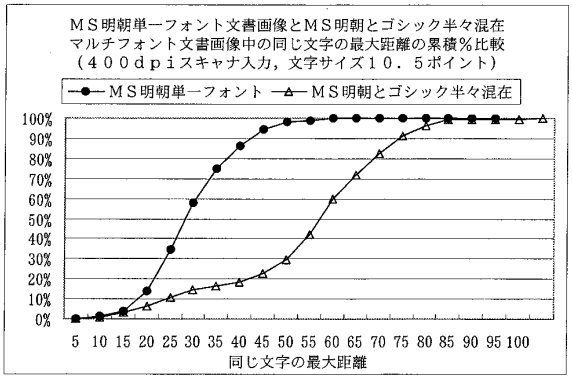


図5.1 文書画像中の同じ文字の最大距離の累積%の比較

平均適合率を求めた結果である．図5.2は，表5.1中の画素数と平均再現率および平均適合率の関係をグラフで示したものである．画素数が上がると平均再現率および平均適合率ともに上がっていくことがわかる．特に，53×50画素（400dpiの10.5ポイント）以上では，平均再現率が99.7%のときに平均適合率が85%を越える高い値になっている．また，1文字の画素数が45×42以上（辺の分割数は8）の場合は，それよりも小さい画素数（辺の分割数6）に比べて平均再現率が高くなっているにも関わらず平均適合率が10〜20ポイント以上も高くなる．この結果は，辺を8分割すると異なったフォントの形状に対してもロバストな特徴をうまく抽出できることを示している．目黒・梅田は，文字サイズが写植32級（8 mm）の場合に辺を8分割にすると文字の分類率が極めて高い値になることを示している（目黒・梅田，1982）が，この結果と合致している．以上のことから，マルチフォント文書画像で高い検索精度を得るためには，最低でも45×42画素程度の画素数を確保する必要があることが分かる．一方，MS明朝単一フォントだけの場合についてみると，表5.2に示す通り解像度を上げると平均再現率100%時の平均適合率も上がるが，マルチフォントの場合ほど画素数による違いは大きくはない．これは，単一フォント文書画像の場合は同じ文字の形状の違いがMS明朝とMSゴシックの場合に比べて非常に小さく，200dpiという低い解像度でも高い検索精度が得られているためである．

表5.3は10.5ポイント文字のMS明朝とMSゴシックを1行おきに混在させた400dpiのマルチフォント文書画像の1,628個の検索文字（平均出現頻度17.05）の平均再現率と平均適合率および非適合係数を詳細に示したものである．許容距離＝81のときは，平均再現率が99.53%，平均適合率は87.11

表 5.1 MS 明朝と MS ゴシックを 1 行おきに混在させたマルチフォント文書画像の解像度と文字サイズ別検索結果

文字サイズ	画素数 (dpi)	辺の分割数	許容距離*	53/74	55/76	57/78	59/80	61/82
10.5p	26×24 (200)	6 分割	平均再現率 (%)	90.45	92.57	94.40	95.86	97.09
			平均適合率 (%)	95.80	92.97	89.18	82.06	73.86
10.5p	39×37 (300)	6 分割	平均再現率 (%)	93.60	95.07	96.55	97.68	98.55
			平均適合率 (%)	95.26	92.53	88.29	82.58	75.49
12p	45×42 (300)	8 分割	平均再現率 (%)	95.72	96.63	97.60	98.30	98.88
			平均適合率 (%)	95.66	93.51	90.83	87.15	82.99
10.5p	53×50 (400)	8 分割	平均再現率 (%)	97.03	97.98	98.80	99.30	99.70
			平均適合率 (%)	96.16	94.49	92.06	89.06	85.18
12p	61×57 (400)	8 分割	平均再現率 (%)	99.25	99.44	99.52	99.62	99.79
			平均適合率 (%)	96.28	94.78	92.83	89.88	85.92

* 6 分割時の許容距離／8 分割時の許容距離

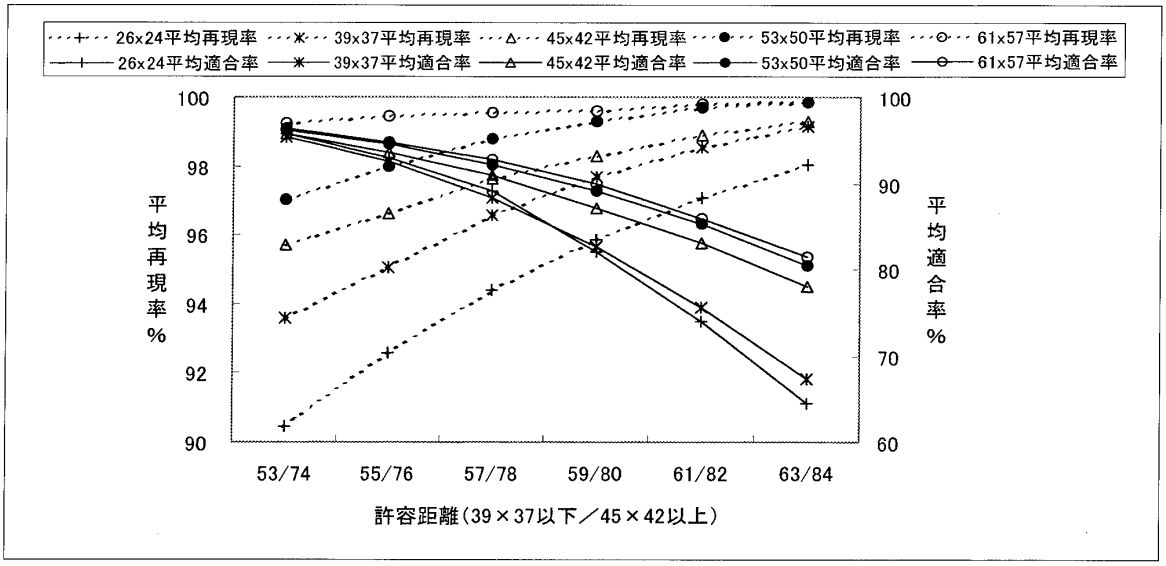


図 5.2 MS 明朝と MS ゴシックを 1 行おきに混在させたマルチフォント文書画像の画素数と検索精度の比較

表 5.2 10.5 ポイント MS 明朝単一フォント文書画像の解像度別検索結果

文字サイズ	画素数 (dpi)	辺の分割数	許容距離*	50/70	51/71	52/72	53/73	54/74
10.5p	26×24 (200)	6 分割	平均再現率 (%)	99.98	99.98	99.99	100	100
			平均適合率 (%)	95.52	94.20	92.76	90.95	88.56
10.5p	39×37 (300)	6 分割	平均再現率 (%)	99.97	99.99	100	100	100
			平均適合率 (%)	96.98	96.22	95.39	94.41	92.94
10.5p	53×50 (400)	8 分割	平均再現率 (%)	99.99	99.99	100	100	100
			平均適合率 (%)	98.31	97.81	97.17	96.51	95.64

* 6 分割時の許容距離／8 分割時の許容距離

%，非適合係数は0.129である．許容距離＝82のときは，平均再現率が99.7%，平均適合率85.18%，非適合係数は0.148である．ところで，許容距離＝81と82の時に再現率が100%にならない検索文字列の個数はそれぞれ全体の約3.38%（55個），2.3%（38個）と極めて少ない．許容距離＝82で再現率100%にならない文字は最大距離が82を越えるもので，図5.3に示す〝温〟，〝示〟，〝圧〟，〝互〟，〝形〟，〝係〟を含む文字であった．これはプリンタの出力精度にも関係してくるが，いずれも文字周辺のはねの位置や直線のおさえ部分のノイズや歪みによりMSゴシックとMS明朝の外郭辺の切り出しに若干違いが認められる．しかし，図5.4のように形状の違いが目で見て分かるものでも最大距離が60以下と小さいものもある．これは，第1，第2ペリフェラル特徴に大きな影響を与えないような内部的な違いにとどまっていることによると思われる．一般にペリフェラル特徴は，内部の潰れや歪みには比較的強いが，周辺のノイズや歪みの影響を受けやすい．これは文字画像の外郭辺の切り出しに影響を与えるからである．以上のことから，許容距離＝82のときには再現率が100%に満たない文字が全体の2%程度あるが，MS明朝とMSゴシックが半々程度混在するという厳しい条件下での実験結果であることを考慮すると，これはやむを得ない範囲と考えてもよいと思われる．また，このとき非適合係数が0.148であるということは，10個程度の2文字漢字がヒットしたときに多くても2個に満たない数の非適合文字が検索されるという程度なので，十分実用に耐え得る精度を示していると言ってよい．

以上の結果，通常の文書でよく使用される10.5ポイント文字で印刷されたMS明朝とMSゴシックを1行おきに混在させたマルチフォント文書の場合は，400dpiの解像度でスキャナ入力すると実用上問題のない検索精度が得られることが確かめられた．

表5.3　10.5ポイントMS明朝とMSゴシックを1行おきに混在させた400dpiのマルチフォント文書画像の1,628組の2文字漢字検索結果

許容距離	80	81	82	83	84
平均再現率（%）	99.30	99.53	99.70	99.77	99.84
平均適合率（%）	89.06	87.11	85.18	82.97	80.48
非適合係数	0.109	0.129	0.148	0.170	0.195
再現率が100%未満の2文字組の個数	65	55	38	30	22
再現率が100%未満の2文字組の割合(%)	3.99	3.38	2.3	1.84	1.35

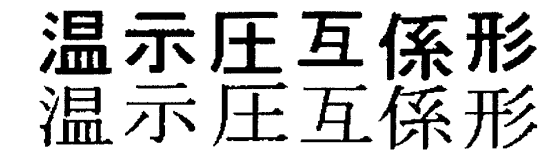


図5.3　許容距離＝82で再現率100%にならない文字

図5.4　形状が違っても最大距離が比較的小さな文字

5.2　従来方式との比較

ここでは，トランスメディアシステムにおいて従来から研究されてきた特徴量（遊佐・田中，1994）を適用した場合と比較することにする．従来方式の特徴量の求め方は次の通りである．まず，文字を覆う最小の矩形領域を切り出し，この文字領域を図5.5のような領域に分割する．次に，各々の部分領域内の黒画素の密度を計算し，分割領域のうち互いに隣り合い，しかも組み合わせると正方形となる分割領域の間で比をとり，これを特徴量とする．その結果18個の特徴量が求められる．これを用いて，2．で示した方法で擬似コードを生成し，文字列検索実験を行った．その結果，MS明朝とMSゴシックが1行お

きに混在するマルチフォント文書の場合は，図5.6に示す通り解像度を上げると平均再現率は上昇するが，400dpiの場合でも表5.3の筆者等の方式に比べると極めて低い精度になっている．このことは，導入する特徴量により検索精度が大きく影響を受けることを示している．以上のことから，筆者等が提案する特徴量の方がはるかに効果的であることが確かめられた．

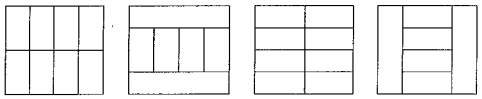


図5.5　トランスメディアの従来方式の文字領域分割パターン

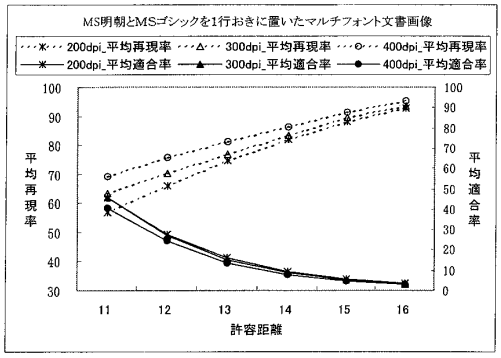


図5.6　マルチフォント文書の従来方式の解像度別精度比較

6．まとめ

トランスメディアシステムの日本語活字文書画像を対象とする文字列検索手法において，文字の形状特性のみに注目した文字列検索手法のフォントロバスト性を検証した．評価実験では，総文字数約4万字の科学分野の抄録集を用いて，10.5および12ポイントのMS明朝とMSゴシックを1行置きに混在させたマルチフォント文書を作成し，200，300，400dpiでスキャナ入力した文書画像を検索対象として用いた．検索精度の評価は，文書画像の最初の約1万字を対象に平均再現率，平均適合率，非適合係数（＝Σ非適合件数/Σヒット件数）を求めて行った．

以上の結果，次のことが明らかになった．

1）解像度（画素数）を上げると検索精度が上がる．

2）400dpiの10.5ポイントのマルチフォント文書画像を対象に1,628組の2文字漢字（平均出現頻度17.05）で検索した結果，許容距離＝82のとき，平均再現率が99.7%で平均適合率85.18%という高い検索精度が得られる．このとき，非適合係数は0.148で実用上問題にはならない程小さく，再現率が100%未満の2文字漢字も約2.3%と僅かであった．

3）通常よく使用される文字サイズ10.5ポイントのマルチフォント文書画像の場合には，400dpi相当で実用に耐え得る高い検索精度が得られる．

一般的な文書では，タイトルや強調文字などにしかゴシック体は使用されない．タイトルには重要語が多く含まれるので，文書中に同じ文字が出現する確率は高くなる．しかし，実験で使用した抄録集のタイトルを全てMSゴシックにしても，MSゴシックとMS明朝の両方を有する文書中の2文字漢字は，実験で用いた2文字漢字全体の30%程度である．従って，2）よりも検索精度は上がっても下がることはない．また，現在，スキャナーの解像度は飛躍的に上がっているが，筆者等の提案する検索手法では，400dpi程度の解像度でも十分実用に耐え得る高い検索精度が確保できることが示された．

以上，筆者等が提案する文字列検索手法は，マルチフォント文書画像に対しても実用上問題のないフォントロバスト性を有するものであると言ってよい．

ところで，予め文書画像毎に，疑似コードと共に擬似コード化に使用したヒストグラムを蓄積しておくと，検索文字列を検索対象の各文書画像のヒストグラムを用いて擬似コード化することで，複数の文書画像を横断的に検索することも可能になる．また，本手法で

は，文字認識を行わないため，辞書が存在しない言語などに対しても適用可能となり，応用範囲は広いと考えられる．

なお，システム化に際しては最適な許容距離のチューニング方法を検討する必要がある．これは，今後の課題である．また，ここで用いた科学技術抄録集とは文書の趣が異なる文学作品や評論，用語集などにも適用して，本手法の分野独立性を示すこと，更に，分かち書きされていない言語や比較的印刷文字に近い楷書で書かれた手書き文字などに本手法を適用することも次の課題である．

謝辞

本研究は本学部の理系教員研究プロジェクトの助成金を受けて行われました．記して感謝します．

参考文献

梅田三千雄（1979）「マルチフォント印刷漢字の分類」，信学論（D），62－D，2，pp.133－140.

太田学，高須淳宏，安達淳（1998）「認識誤りを含む和文テキストにおける全文検索手法」，情報処理学会論文誌 Vol.39，No.3，pp.625－635

田中知朗，田中譲（1997）「トランスメディアシステムによる英文テキスト画像処理」，情報処理学会論文誌 Vol.38，No.7，pp.1389－1398

新國三千代，田中譲（2001）「10.5 ポイントで印刷された日本語文書画像の文字列検索―トランスメディアシステムにおける検索手法の改良―」，札幌学院大学社会情報学部紀要『社会情報』，Vol.10，No.2，pp.31－45.

丸川勝美，藤澤浩道，嶋好博（1995）「文字認識と全文検索の融合技術に関する実験的検討」，情報処理学会研究報告，95－FI－39，pp.65－72

目黒眞一，梅田三千雄（1982）「マルチフォント印刷漢字の認識」，信学論（D），65－D，pp.1026－1033.

遊佐実，田中譲（1994）「トランスメディアシステムの日本語への拡張」，第 49 回電子情報通信学会全国大会論文集，2H－9，pp.217－218.

遊佐実，田中譲（1995）「画像文書に対する多言語文字列検索機能の実現」北海道大学工学部電気工学専攻応用制御講座修士論文

2002 年 1 月 30 日受付
2002 年 2 月 15 日受理