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——

新國三千代・田中 譲

This paper proposes a full-text search method for Japanese document images. The method improves the one used by the "Transmedia System" that performs a full-text search of document images without using OCR technologies. Our new method encodes each Japanese character image to a 'pseudo code', using the "Peripheral Features". The method is capable of searching document images for an arbitrary specified character string. This search is essentially based on the well-known string-matching algorithm. It achieves both the 100% average recall ratio and the 88.7% average precision ratio in the search for 1,750 different two-Kanji-character terms. This experiment used 200dpi document image including about 10,000 Japanese characters of 10.5 point MS-Minchou font.

1．はじめに

15世紀の半ばにグーディンベルグにより印刷技術が発明されて以来，世界中で膨大な量の印刷文書が蓄積されているが，これらを機械可読化して蓄積することは不可能と言われている．一方，このような文書をカメラ撮影やスキャナーなどで画像化することは比較的容易に行えることから，文書をそのまま画像化してデータベース化する試みが進行している（このように画像化された文書を文書画像と呼ぶ）．しかしながら，これらの膨大な文書画像から必要な情報を引き出すことは難しく，文書画像データベースにおける課題の一つになっている．

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

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．いずれの場合も再現率を100%にすることは難しく，実現はされていない．

一方，筆者らは，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（文字画像と呼ぶ）を何らかの特徴量で擬似コード化し，この擬似コードを用いて文字検索を行うものである．この方

法では，再現率を100%にすることが可能である．また，OCRのように辞書を用いて文字認識を行う必要がないため，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．

今までのトランスメディアシステムの研究において，遊佐らは日本語の文字を形作る黒画素の密度比を特徴量として用いて擬似コードを生成し，文字を検索する方法を提案している．ここでは24ポイントという大きな印刷文字を対象に精度評価を行っている．これを日本語の印刷文字で通常よく使用されている10.5ポイント（活字では5号サイズ）の文書画像に適用してみると，あまりよい精度が得られない．そこで，筆者は，文字の形状の周辺情報に基づくペリフェラル特徴（梅田，1979）に注目し，これを用いて擬似コードを生成することを考えた．ペリフェラル特徴は文字フォントにロバストな特徴量として提案されたものであり，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こともこれを採用した理由の一つである．本研究では，この擬似コードを用いて文字画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を検索する方法を提案する．この方法によると，再現率100%でしかも十分実用に耐えうる適合率を得ることができる．

一般に印刷文書には種々の文字フォントが混在している場合が多いが，本稿では，対象を10.5ポイントのMS明朝で印刷された日本語文書に限定して議論を進めることにする．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，低い解像度でもここで提案する方法が有効であることを示すために敢えて200 dpiという低解像度の文書画像を対象に検証実験を行っている．

以下，2．で従来のトランスメディアシステムにおける文字画像の擬似コードの生成と検索手法，3．で筆者が提案する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，4．で文字画像の距離の定義と検索手法，5．で実験の準備と実験方法，6．で実験結果と考察，7．でまとめと今後の課題について述べる．

2．従来のトランスメディアシステムにおける文字画像の擬似コードの生成と検索手法

トランスメディアシステムでは，次のような方法で文字画像から擬似コードを生成する．まず，文字を覆う最小の矩形領域を切り出し（これを文字領域と呼ぶ），この文字領域を複数に分割する，そして，各々の部分領域の面積に対する黒画素の密度を計算し，得られた密度の中から任意に2つを取り出して比をとったものを1つの特徴量とする．図2.1で示す分割領域のうち，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るのが最適であることが実験により示されている．この場合，1つの文字画像につき18個の特徴量が生成される．これらの特徴量を文書画像中のすべての文字画像について求め，各特徴量について横軸が密度比，縦軸が文字数であるヒストグラムを作成する．

次に，図2.2のようにこのヒストグラムの面積がほぼ同じになるように4等分し，該当する特徴量の値がどの領域に含まれるかにより疑似コードを生成する．4個の領域から異

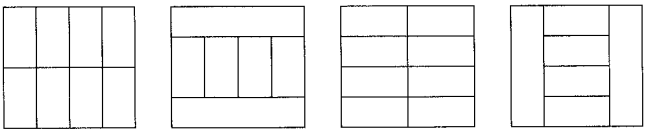


図2.1 従来方式の文字領域の分割パターン

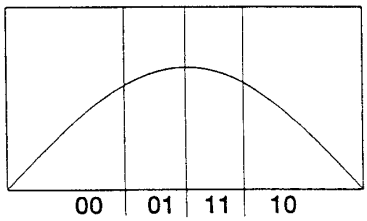


図2.2 ミラーコーディング

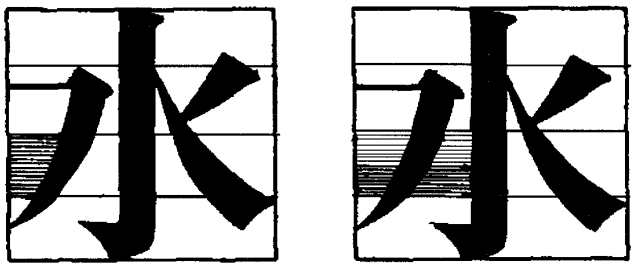
なるコードを生成するためには2ビットあればよいので，各領域を左から00，01，11，10という2ビットコードに対応させる．このとき，互いに隣り合う領域に属する文字は1ビットの違いしか持たないようになっている．このコーディングの方法をミラーコーディングと呼ぶ．ミラーコーディングでは，注目する文字の特徴量が2つの隣り合う分割領域の境界付近に存在し，どちらの領域に属するかが曖昧な場合，仮に誤ってコードが生成されたとしても2ビット中の1ビットしか違わないようすることができる．このようにして求めた2ビットコードを特徴量の数だけ横に並べて1つの擬似コードを生成する．ここでは18個の特徴量を用いているので，1つの文字画像から36ビット（2ビット×18）の長さをもつ擬似コードが生成されることになる．

検索時には，検索文字の各特徴量と照合する文字の擬似コードを1ビットずつ比較していき，異なるビット数がある許容範囲に収まれば同じ文字と見なすことにより文字を検索する．

3．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

3.1 ペリフェラル特徴の導入

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合うまでの文字部でない領域の面積を計数して（白画素を数える），これを当該分割領域の面積で割って規格化する〟（梅田，1979）ことにより抽出される．その結果，得られるペリフェラル特徴は0～1の値となる．図3.1のa）に辺を4分割して左辺からペリフェラル特徴を求める例が示されている．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を反映させるために，図3.1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，二次ペリフェラル特徴を求める．4分割の場合は一次と二次を合わせると32個のペリフェラル特徴が抽出される．梅田氏の実験では，文字サイズが写植32級（8 mm）の印刷漢字で明朝体の12個のフォント，ゴシック体6フォントに対して検証を行い，8分割で一次および二次ペリフェラル特徴を用い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．本研究で扱う印刷文書の文字の大きさは，活字の5号



a）一次ペリフェラル特徴　b）二次ペリフェラル特徴

図3.1 ペリフェラル特徴

サイズ（約3.7 mm，10.5ポイント）でほぼ半分の大きさである．従って，梅田氏の方法を適用しても同じような精度が得られることは保証されていないが，擬似コードを生成する特徴量としてこのペリフェラル特徴を導入することにした．

以下，文書画像からの文字領域の切り出し方法，文字領域の分割，ペリフェラル特徴の傾向について述べることにする．

3.1 文書画像からの文字領域の切り出し

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．各頁は水平方向および垂直方向に傾きがないように置かれているものとする．

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）とし，横をｘ座標，下方にｙ座標をとる．

2）行を切り出す：水平方向に走査し，初めて黒画素が出現する位置のｙ座標を行の上端とする．次に，垂直方向に座標を1つずつ動かし，水平方向に白画素のみが継続して出現した位置のｙ座標－1を行の下端とする．

3）文字領域の左右外郭辺(文字の幅に相当)を切り出す：行の左上端から水平方向にｘ座標を1つずつ動かしながら垂直方向に行の下端まで走査し，垂直方向に始めて黒画素が出現する位置のｘ座標を文字領域の左辺のｘ座標とする．さらにこれを続けて，垂直方向に始めて白画素のみが出現する位置のｘ座標－1を文字領域の右辺のｘ座標とする．

4）文字領域の上下外郭辺（文字の高さに相当）を切り出す：3）で切り出された左外郭辺の左上端から垂直方向下にｙ座標を1つずつ動かしながら，水平方向に右外郭辺まで走査し，水平方向に始めて黒画素が出現した位置のｙ座標を文字領域の上辺のｙ座標とする．同様にして，左外郭辺の左下端から垂直方向上にｙ座標を1つずつ動かしながら，水平方向に始めて黒画素が出現した位置のｙ座標を文字領域の下辺のｙ座標とする．「っ」と「つ」のように高さが違うが形状が同じという文字については，文字の高さがその行の高さの1/2以下の場合は平均の高さを持つ文字の上辺および下辺のｙ座標とした．これにより「っ」と「つ」などを異なる文字と見なすことができる．また，「一」「ー」のように極端に文字の高さが低いものについても，ノイズによって高さの影響が大きく出るため，便宜上同様に高さを補正した．

1）～4）の結果，文字領域の大きさは文字により異なってくる．また，3）の処理では「い」や「は」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2文字となる．全ての同じ文字が同様に分離されていれば問題はないが，ノイズの入り方により同じ文字が分離されたりされなかったりする場合がある．実際，本研究の実験に用いた文書では，このように分離された文字は52字出現する．これについては，文字領域の幅が平均の長さに満たない場合は，後続する次の文字領域と合わせて一つの文字領域と見なすことにより対応することもできる．しかし，ここで用いる文書画像ではこのような文字が全体の5％と少ないことと，文書画像の領域切り出しの研究は既に行われている（秋山等，1984）ことから，本研究では特にこの問題は取り扱わないことにした．従って，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．

3.2 文字領域の分割

本研究の実験で使用する約1万字のMS明朝の文書画像の文字領域の高さと幅の平均を求めると，10.5ポイントの文字で解像度が200 dpiの場合，平均の高さは約22画素，平均の幅は分離文字を除くと約25画素，分離文字や英数字を含めると平均の幅は約18画素と小さい値になる．今仮に，文字領域の外郭辺

をそれぞれ4分割してペリフェラル特徴を求めることにすると，分割の際，1）辺の長さが4で割り切れない，2）分割数が4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，といった問題が生じる．この場合は，次のようにして分割幅を決めることにする．1）では，小数点以下切り捨てた値を分割幅とする場合と四捨五入をする場合があるが，ここでは切り捨てた値を分割幅とする（実験の結果では，どちらをとっても検索精度に大きな違いは出ない）．2）では，満たない分の値を0とする．3）は，4番目の分割幅が他と異なる場合である．この場合は，4番目の分割幅は残り幅とした．例えば26を4分割すると1）から分割幅が6，6，6，8となる．

3.3 ペリフェラル特徴の傾向

200 dpiでスキャナー入力した文字はノイズが介在するため同じ文字が全て同じパターンになることは極めて稀である．実際，文書画像中の各文字についてペリフェラル特徴を求めると，特殊な記号を除くと同じ文字が同じ値を持つことはない．図3.2は文書中に出現する21個の〝階〟の文字領域の各辺を4分割した一次ペリフェラル特徴のグラフである．X軸は，上・右・下・左辺の順にそれぞれ4分割した領域を左辺または上辺から順に1～16のラベルをつけて並べている．全体的な傾向は似ているように見えるが同じものはないことが分かる．図3.3は各分割領域毎の一次ペリフェラル特徴の最大値と最小値のグラフである．一般に，この値は印刷やスキャナー入力時のノイズの入り具合で変わってくる．

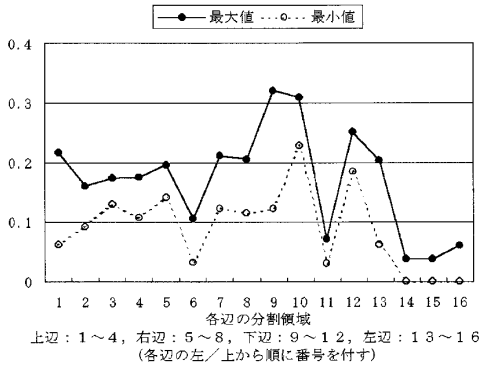


図3.3 21個の〝階〟の4分割一次ペリフェラル特徴の最大値と最小値

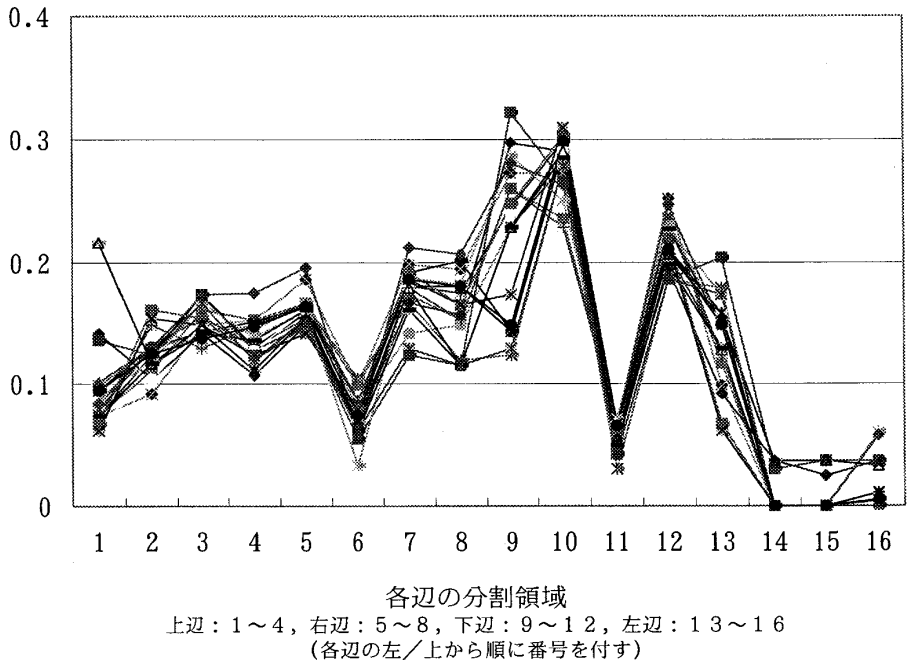


図3.2 21個の〝階〟の4分割一次ペリフェラル特徴

3.4　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

本研究では，各分割領域毎に2．で述べたヒストグラムを作成する．各辺4分割の場合は，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個のヒストグラムを作成することになる．図3.4に示す通り，ヒストグラムの形状は分割領域によって異なっている．ヒストグラムを作成する際のクラスは，0～1の値をとるペリフェラル特徴を0～255に対応させて256個で求めている．これ以上細かくクラスを分けても生成されるコードがほぼ同じ値となるためである．

疑似コード化は，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等分して，ヒストグラムの各部分領域に含まれる文字毎に2^nビットコード（n＝2の場合は，00，01，11，10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）を生成する．以下，この各コードを単位コードと呼ぶことにする．1つの文字画像の疑似コードは，単位コードを特徴量の数だけ横に並べたものなので，　例えば，各辺4分割で2ビットコードを生成すると，一次ペリフェラル特徴から生成される疑似コードは16特徴×2ビット＝32ビットの長さを持つ疑似コードとなる．二次ペリフェラル特徴も含めると，32特徴×2ビット＝64ビットの長さを持つ疑似コードが生成されることになる．

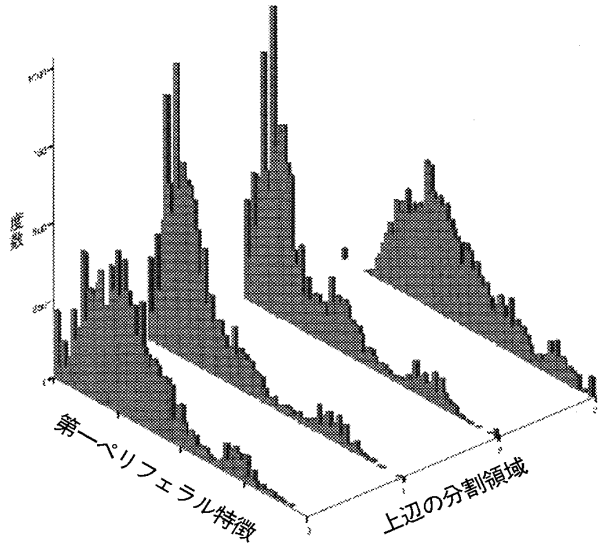


図3.4　上辺4分割の分割領域別ヒストグラム

4．文字画像の距離の定義と検索手法

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て，距離的に近い文字，すなわち，検索文字との距離が指定された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．

ここでは，検索文字に指定された1文字画像Aと検索対象の文書画像中の照合される1文字画像Bとの間の距離を下記のように定義する．

AとBの第i番目の単位コードを，a_i，b_i，

$$c_i \equiv \left\{ \begin{array}{l} 0：a_i と\ b_i が\ i\ 番目の分割領域のヒストグラムで同じ部分領域に含まれる \\ k：a_i と\ b_i が\ i\ 番目の分割領域のヒストグラムで\ k\ 個離れた部分領域に含まれる \\ （隣り合う場合を1とする，k＝1～2^n-1，nは単位コードのビット数） \end{array} \right\}$$

とするとき，

AとBの距離d_{AB}を

$$d_{AB}=\sum_{i=1}^{m} c_i \quad (m＝ペリフェラル特徴の数)$$

と定義する．

これは，疑似コードを生成するヒストグラムにおいて，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ように重みをつけていることを意味している．一般に，スキャナーで入力した文書画像にはノイズが入っているため，AとBが同じ文字であってもその特徴コードが全て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ことにより文字を検索する．検索文字で2文字以上を指定した場合は，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，これを同じ文字と見なすことにより文字列を検索する．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．

検索の精度は，下記のように検索文字に対する再現率(recall ratio)と適合率(precision ratio) で与えることにする．

再現率＝正しくヒットした数/検索されるべき対象文字の数×100

適合率＝正しくヒットした数/ヒット数(誤って抽出された文字を含む)×100

検索時に検索文字を指定する方法は，文書画像中の文字画像すなわち疑似コードを直接指定するか，キーボードからMS明朝で入力した文字画像を疑似コード化したものを用いて行う．後者の場合，入力した文字の疑似コード化は検索対象の文書画像で作成したヒストグラムを用いて行う．

図4.1は文書画像中の〝星〟という文字を検索文字に指定して，照合される他の文字との距離を分布で示したものである．この例は，総文字数が約4万弱の文書を対象に一次ペリフェラル特徴の16個のヒストグラムを用いて32ビットの疑似コードを生成し，最初の1,568文字(文書画像の1頁) 中に出現する2つの〝星〟(図中星A，星B) をそれぞれ検索文字に指定して，1頁に含まれる他の全ての文字との距離を求めたものである．1頁には

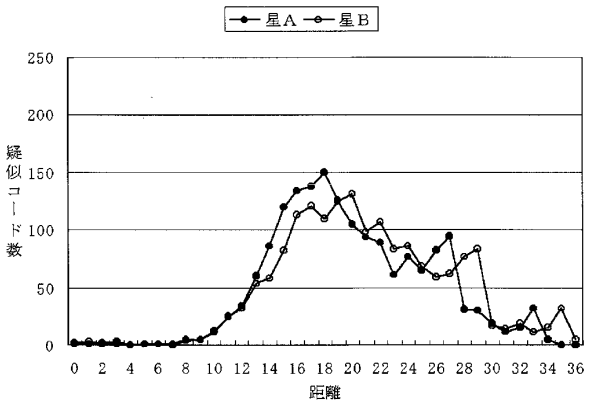


図4.1 文書画像1頁中の2つの〝星〟(星A，星B) と他の文字との距離分布

8個の〝星〟が出現するが，検索文字の星Aおよび星Bと1頁内の他の〝星〟との最大距離は5，検索文字の〝星〟とは異なる文字の最小距離は6である．従って，この場合に許容距離＝5とすると，適合率，再現率とも100%になる．実験によると，一般に距離が小さい部分に分布する疑似コードの数は全体の中で極めて少なく，あるところからグラフが急激に立ち上がり右側になだらかな尾をひく形になっている．

5．実験の準備と実験方法

ここでは，まず実験に使用する文書画像について述べる．次に，信頼性の高い実験を行うために，使用する文書画像の文字数と疑似コードの関係について調べ，実験の方法と評価尺度の求め方について述べる．

5.1 実験に使用した文書画像

実験に用いる文書として，天文学会のホームページにPS(ポストスクリプト) の形で公開されている年会講演の抄録集を用いた．ここからはテキスト情報も得られるため実験の評価が容易に行えるためである．因みに，抄録集のタイトルは14ポイントのゴシック体で本文やその他の文字は10.5ポイントのリュウミンライトで書かれている．

本研究の実験では，1999年度の春季学会の「星形成」分野の44件の抄録(分野，キーワー

ド，タイトル，概要からなる）を用いた．実験で用いた抄録には英数字や記号も含まれる．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（異なる文字の数）で総文字数は38,566字である．A4版で1頁に1行40字×50行でべた打ちすると22頁になる．図5.1は，1頁目から各頁までの総文字数と異なる文字数，総文字数中の異なる文字の割合を示している．異なる文字数と総文字数の関係は，2万字～3万字の学術論文などに見られる傾向と殆ど同じである．

この文書をワープロに読み込み，文字フォントをMS明朝，文字サイズを10.5ポイントにして600 dpiで印刷し，それを200 dpiでスキャナー入力して2値モノクロの文書画像を作成する．本方式ではスキャナーの解像度だけではなく印刷時のノイズの影響も無視できないが，このようなノイズに対しても堅牢な手法が提案できればよいと考えこれを容認した．

なお，本実験で用いたシステムはJAVAで作成されており扱える画像形式に制限があることから，2値モノクロの画像データを24ビットのjpegに変換したものを利用している．従って，実際には2値モノクロに比べ更に精度が落ちたデータとなっているが，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，モノクロと変わらない文字パターンが得られるためこのことが精度に大きな影響を与えることはない．

図5.2はこのようにして得られたMS明朝の〝層〟と〝形〟という文字画像を拡大したものである．微細なところで文字の形に違いがあることが分かる．

5.2 疑似コードと文書画像の総文字数との関係

信頼性のあるデータを取得するためには，どの程度の文字数の文書を用意すればよいのかを予め確認しておく必要がある．そこで，

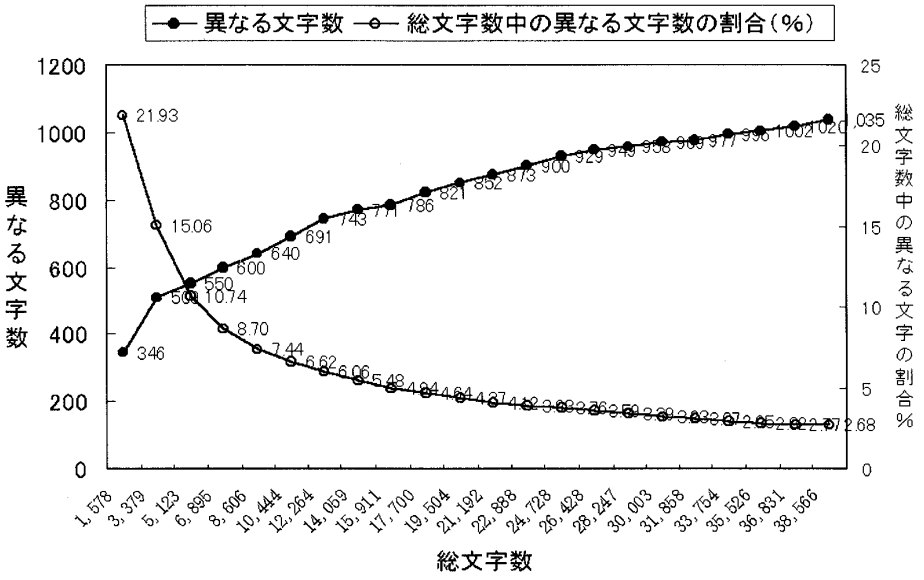


図5.1 実験に用いた文書画像の総文字数と異なる文字数の分布

図5.2 10.5ポイントのMS明朝の文字画像例

疑似コードを生成するのに際し，文書画像の総文字数がどのように影響するのかを調べるために，次のような実験を行った．まず，5.1で述べた文書画像を，〈A〉文書画像の1頁(総文字数1,578字，異なる文字数346字)，〈B〉文書画像の1～3頁(総文字数5,123字，異なる文字数550字)，〈C〉文書画像の1～6頁(総文字数10,444文字，異なる文字数691字)，〈D〉文書画像の1～22頁(総文字数38,566文字，異なる文字数1,035字)の4つに分けて作成する．次に，1頁目の文書画像の1,578文字について，〈D〉で生成された疑似コードと〈A〉～〈C〉で生成された同じ位置にある文字の疑似コードとの距離(4．の定義と同様)を求め，総文字数との関係を調べる．疑似コードを求めるヒストグラムの分布が文字数に依存しない場合は，同じ位置にある同じ文字同士の距離は0になるはずである．図5.3は文字領域の外郭辺を4分割して16個の一次ペリフェラル特徴を用いて生成された疑似コードの実験結果である．グラフ上〈B〉，〈C〉は同じ傾向を示している．〈B〉では，距離0が67.5%，距離1が25.9%，〈C〉では，距離0が77.6%，距離1が19.3%で距離0と1を合わせると，〈B〉93.4%，〈C〉96.9%となる．このことから，本実験で用いている文書の場合は，総文字数約5,000字以上あればほぼ同じ疑似コードを生成することが分かる．しかし，実験では安定したデータを取得するために〈D〉の文書画像を用いて疑似コードを生成することにした．

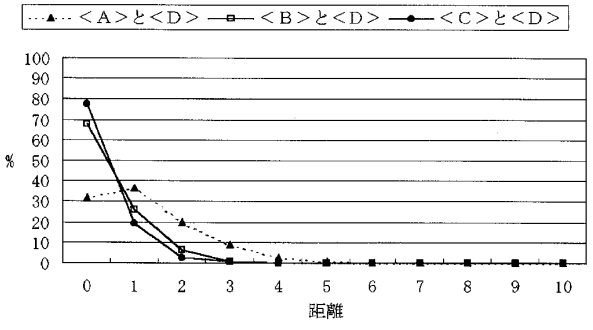


図5.3 擬似コードと文書画像の文字数の関係

5.3 実験の方法と評価尺度

5.1で述べた22頁の文書画像中の4万弱の全文字画像と生成された疑似コードの対応関係を付けることは，印刷汚れによるノイズや分離文字が介在することから大変な作業となる．そこで，最初の6頁分の約1万文字(総文字数が10,444字，異なる文字数691字)に対して検索の結果を検証することにした．5.2の結果からも全体を概観するにはこれで十分である．この1万文字について疑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．

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，分離文字と出現頻度1のものを除いて約1,750個の2文字漢字を検索用文字列として作成する(ここで言う検索用文字列とは2文字画像の擬似コードが並んだものである)．

2) 1)で抽出した1,750個の2文字漢字を検索文字に指定して1万弱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，検索の成否を判定するプログラムを実行して再現率，適合率を求める．

4) 1,750個の2文字漢字の平均再現率と平均適合率を求めて，検索精度の尺度とする．

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ことから1)のような方法をとった．また，3文字以上については，検索文字と同じ並びの文字列が出現する確率は2文字の場合よりもさらに少なくなるため精度は上がることになる．従って，2文字の場合について精度を評価しておけば十分である．

本実験では，文書中の文字画像で検索用文字列を生成しているが，実際には2文字からなる検索用文字列をキー入力し，MS明朝の文字画像に変換して，これを用いて検索を行

うことが多いと思われる．この場合は，検索文字画像を擬似コード化する際に，検索する対象の文書画像のヒストグラムを用いることにより，特別なノイズが介在しない限り，ある距離範囲内に収まるような擬似コードが作成される．

6．評価実験の結果と考察

検索精度の評価を行う前に次のことを調べておく必要がある．

a）使用する特徴量は一次ペリフェラル特徴のみでよいのか，一次と二次ペリフェラル特徴の2つを用いた方がよいのか，

b）文字領域の辺の分割数はいくつが適当か，

c）単位コードのビット数は何ビットが適当か，

これらを明らかにするために採用するペリフェラル特徴と辺の分割数そして単位コードのビット数をパラメータにして下記の実験を行った．これらのパラメータは相互に関連し合っているので，適宜これらを組み合わせて実験を行った．

1）採用するペリフェラル特徴の検証

2）最適な辺の分割数の検証

3）単位コードの最適なビット数の検証

以下，6.1～6.3で1）～3）の実験結果と考察を述べ，6.4で従来の方式との比較を行い，6.5でここで提案する検索手法の精度評価について考察する．

6.1 採用するペリフェラル特徴の検証

一次ペリフェラル特徴だけを用いて，辺の分割数を4，6，8にして2ビットおよび3ビットコードで実験を行った．その結果，6分割で3ビットコードが最も精度が高く，平均再現率100%時の平均適合率は74.09%である．他の場合は，平均再現率100%時の平均適合率は20.58%～51.53%と低い値になっている．一次ペリフェラル特徴は文字の外側の特徴だけを用いているので，内部構造が異なる文字も同じ文字とみなされ，適合率が下がるためである．

図6.1は，上記の実験で最も適合率が高かった6分割3ビットコードで，一次ペリフェラル特徴だけを用いた場合と一次と二次ペリフェラル特徴を用いて実験を行った結果を比較したものである．X軸は，検索時に同じ文字とみなす許容距離を上記の2つの場合について平均再現率が100%になる距離で揃えて並べたものである．一次ペリフェラル特徴だけを用いた場合は距離＝28，一次と二次ペリフェラル特徴を用いた場合は距離＝53で平均再現率が100%になる．距離の定義（$d_{AB}=\sum_{i=1}^{m} c_i$，$m$＝ペリフェラル特徴の数）から，許容距離は特徴の数に比例して大きな値になる．図6.1より，第一と第二ペリフェラル特徴を用いると，一次だけを用いた場合よりも平均再現率100%時の平均適合率が1.2倍高くなることが分かる．その他の場合も，第一と第二ペリフェラル特徴を用いると，一次だけを用いた場合よりも平均再現率100%時の平均適合率が1.7～3.2倍以上も高くなる．このことから，一次と二次のペリフェラル特徴を用いる方が高い精度が得られることが確かめられた．

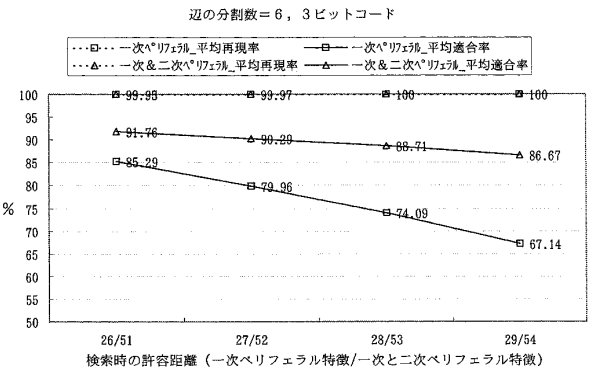


図6.1 採用するペリフェラル特徴と1,750個の漢字2文字の平均再現率および平均適合率

6.2 最適な辺の分割数の検証

辺の分割数を4，6，8として，一次ペリ

フェラル特徴と二次ペリフェラル特徴の二つを用いて実験を行った．

図6.2は，2ビットコードで1,750個の漢字2文字を検索文字に指定して検索した結果である．分割数＝4では距離＝19，分割数＝6では距離＝26，分割数＝8では距離＝38で平均再現率が100%になっている．このときの2ビットコードの平均適合率は，分割数＝4で49%，分割数＝6では77.74%，分割数＝8では59.49%になり，分割数＝6の場合に最も高い平均適合率が得られることが分かる．これは最低の4分割の場合の約1.6倍になっている．

図6.3は，3ビットコードで同様の実験をした結果である．この場合も，辺の分割数＝6が距離＝53で平均再現率100%となり，このとき最も高い平均適合率を示し，最低の8分割の平均適合率の1.5倍になっている．いずれの場合も，4分割では特徴を捉えるのに不十分で，8分割では細分化しすぎて文字の特徴が捕らえられなくなっていることが分かる．

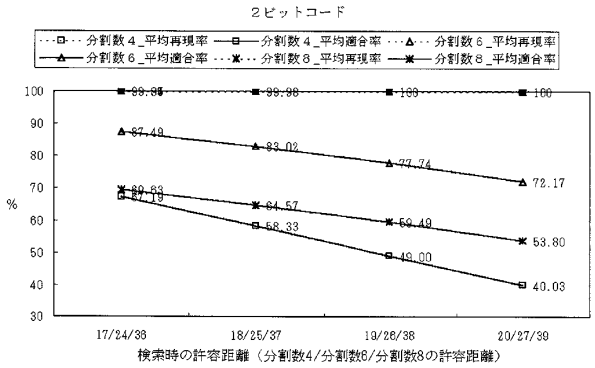


図6.2 辺の分割数と1,750個の漢字2文字の平均再現率および平均適合率（2ビットコード）

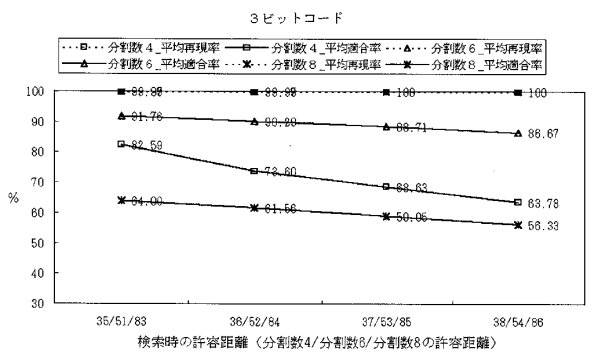


図6.3 辺の分割数と1,750個の漢字2文字の平均再現率および平均適合率（3ビットコード）

以上の結果，200 dpi文書画像の10.5ポイントの文字の場合には，最適な辺の分割数は6であることが確かめられた．

6.3 単位コードの最適なビット数の検証

6.1と6.2の結果を踏まえ，ここでは，一次と二次ペリフェラル特徴を採用し，辺の分割数を6とする．図6.4は，そのときの単位コードのビット数を2，3，4ビットとして実験を行った結果である．ビット数が多いほど平均適合率の傾斜がなだらかになっているが，平均再現率100%で最も平均適合率が高いのは3ビットコードの88.71%で，4ビットコードの87.13%の方が下回っている．4ビットコードで生成される擬似コードのビット数は192であり，3ビットコードの144ビットの1.3倍になるにもかかわらず検索精度が3ビットコードよりも低い結果になっている．単位コードのビット数は，ヒストグラムを何等分するかで決まる値であるが，この値を大きくしたからといって精度を上げることには結びつかないことを示している．以上のことから，ここで対象としている文書画像の場合は3ビットコードが最適であることが分かる．

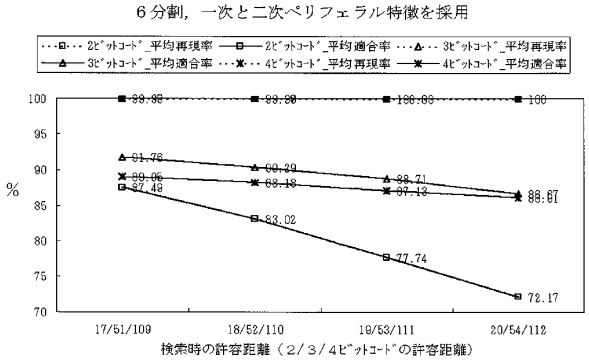


図6.4 単位コードのビット数と1,750個の漢字2文字の平均再現率および平均適合率

6.4 トランスメディアシステムの従来方式との比較

トランスメディアシステムの従来方式は18特徴量を2ビットコードで擬似コード化しているので，1つの文字画像に対し36ビットの長さを持つ擬似コードが生成される．筆者の方法と従来方式を比較するために，擬似コードのビット数の差を出来るだけ少なくして精度の比較を行った．筆者の方式では，精度はよくないが，辺の分割数＝4で一次ペリフェラル特徴のみを用いて2ビットコードを生成すると，擬似コードのビット数が32ビットとなり最も近い値となる．表6.1はその実験結果と従来方式を比較したものである．筆者の方式は，特徴量が従来方式よりも少ないにも関わらず，平均再現率100%のときの平均適合率は従来方式0.18%の約114倍で20.58%になっている．

従来の検索手法は，生成された疑似コードをビット列と見なして照合時には互いに異なるビット数がある範囲内であれば同じ文字と判定している．従って，異なるビット数を距離に対応させることができる．図6.5は図4.1の例で示した，〝星〞（星A，星B）と1頁中の他の文字との距離分布を示したものである．実線が筆者の方式，破線が従来方式である．従来の方式は筆者の方式に比べ狭い凸型になっており，立ち上がり部分が筆者よりも高くなっている．つまり，許容距離を1増やすとそこに入ってくる文字数が増え，その結果適合率が大きく下がることになる．1頁中には8個の〝星〞が出現するが，両方式とも

表6.1 筆者方式と従来方式の比較：1,750個の2文字漢字検索（4分割一次ペリフェラル特徴を採用，2ビットコード）

距離（従来方式／筆者方式）	9/21	10/22	11/23
筆者方式_平均再現率	99.62%	99.83%	100 %
筆者方式_平均適合率	50.34%	33.27%	20.58%
従来方式_平均再現率	99.97%	99.98%	100 %
従来方式_平均適合率	0.24%	0.20%	0.18%

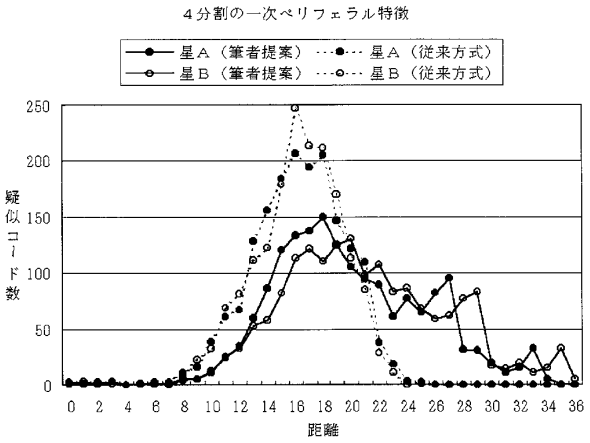


図6.5 MS明朝 星Aと星Bの距離分布（文書画像1頁分）

星Aおよび星Bと1頁内の他の〝星〞との距離の最大値は5，星A，星Bと〝星〞以外の文字との距離の最小値は6になっている．従って，検索時の距離を5とすると両方式とも適合率が100%になるが，仮に距離を8とすると，従来の方式では星A 36.4%，星B 44.4%となり，筆者の方式の星A 53.3%，星B 61.5%に比べ適合率が低くなる．他の文字について比較した結果いずれも従来方式の方が適合率が低くなる．

以上のことから，10.5ポイントの文書画像を検索する手法としては，筆者の方式の方が従来方式よりもはるかに有効である．従来方式は24ポイントという大きな文字について検証されたものであり，10.5ポイントにそのまま適用して議論することは適当ではないが，特徴量を変えるとこのように精度に大きな差が出てくるということが分かる．

6.5 検索精度の評価

6.1～6.3の実験の結果をまとめると，200 dpi文書画像の10.5ポイントMS明朝の文書画像の擬似コード化では，

1）一次および二次ペリフェラル特徴の二つを使用する
2）ペリフェラル特徴を求める際の辺の分割数は6とする
3）ミラーコーディングの際の単位コードのビット数は3ビットとする

と，検索精度が最も高くなることが分かる．

先に示した図6.4の3ビットコードが最も精度が高い場合に相当している．このとき，距離＝53で1,750組の2文字漢字の検索文字の平均再現率が100%となり，平均適合率は88.71%になっている．図6.6は距離＝53の時，すなわち平均再現率が100%になったときの適合率と文字の出現頻度をグラフにしたものである．X軸は，1,750個の文字を出現頻度と適合率の順に並べたものである．この図から適合率が出現頻度に関係していないことがわかる．一般にノイズの影響は，頻度が多いほど大きくなる傾向になると考えられる．しかし，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ことになるので，極めて特異なノイズが出現しなければ，必ずしも出現頻度に反比例して適合率が低くなるわけではない．一方，出現頻度は低くてもよく似た文字が沢山出てくれば，適合率は低くなる．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は少ない．本実験で使用した文書の2文字をとりあげて考えると，頻度が11以上の2文字は全体の4.54%と極めて少ない．10以下では，頻度1が56.47%で最も多く，次は頻度2の20.23%．頻度3～6は合わせて19%である．ところで，検索文字列の頻度が2の場合は，同じ許容距離内に別の文字列が1個でもヒットしてくると，適合率は66.67%以下になる．頻度が10の場合は，適合率は90.91%になる．その意味では，出現頻度が異なる文字列の適合率を同等に議論することはできない．しかし，これを踏まえた上で検索精度を測る目安として適合率を用いることは可能である．図6.7は平均再現率100%時の1,750個の検索文字の適合率の割合を示したグラフである．適合率が100%の2文字漢字の割合は66.63%で，適合率が90%以上のものと合わせると73.20%になる．適合率が40%以下の検索文字は全体の約3%の51組あるが，出現頻度の割合をみると頻度2が56.86%，頻度2～4が96.08%を占める．出現頻度が少ないため，仮に10個違った文字がヒットすると適合率は10数%～30数%と低い値になってしまう．一般に，文字列検索では再現率が100%であれば，関係ない文字が多少ヒットしてきても決定的な問題にはならない．以上のことから，平均再現率100%で平均適合率88.71%という数字は，検索精度としては十分実用に耐えうるものであると言ってよいであろう．

表6.2は，検索時の許容距離と1,750個の検索文字の平均再現率および平均適合率を示したものである．距離＝50では平均再現率が99.95%と100%に達しないものが0.05%あ

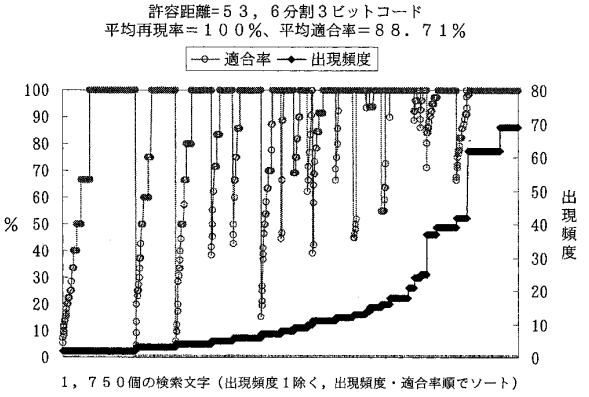


図6.6 1,750個の2文字漢字の平均再現率100%時の適合率と出現頻度

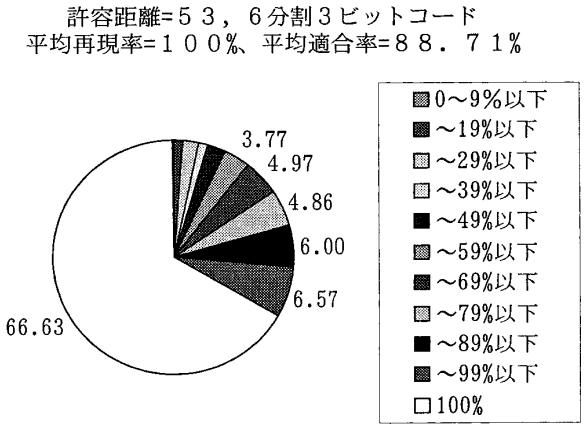


図6.7 1,750個の2文字漢字の平均再現率100%時の適合率の割合

表6.2 1,750個の2文字漢字の検索結果
（6分割，一次・二次ペリフェラル採用，3ビットコード）

距　離	50	51	52	53
平均再現率	99.95%	99.95%	99.97%	100　%
平均適合率	93.16%	91.76%	90.29%	88.71%

る．しかし，これについて調べてみると，出現頻度4の2個の検索文字が1件の検索漏れで再現率75%，出現頻度9の2個の検索文字が1件の検索漏れで再現率88.89%，出現頻度18の2個の検索文字が1件検索漏れで再現率94.44%となっている．出現頻度が少ないと適合率と同様1件の検索漏れが再現率に与える影響が大きくなる．従って，本方式のように高い平均適合率を保ちつつ平均再現率を100%にするということは，かなり厳しい条件であることが分かる．

7．まとめ

文書画像データベースシステムであるトランスメディアシステムにおいて，10.5ポイントで印刷された日本語文書の文書画像を検索する方法を提案した．本方法では，文書画像中の個々の文字画像をペリフェラル特徴を用いて擬似コード化している．そして，この生成された擬似コードに基づき文字画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画像の検索を行っている．

評価実験では，MS明朝の10.5ポイントの総文字数約4万字の日本語文書を200 dpiでスキャナー入力した文書画像を使用して擬似コードを生成している．検索精度の評価は，この文書画像の最初の約1万字を対象に行った．その結果，次のことが明らかになった．

1）辺の分割数を6として一次および二次ペリフェラル特徴を求め，3ビットコードで擬似コードを生成すると最も高い検索精度が得られる．

2）1,750個の漢字2文字を検索文字に指定して再現率，適合率を求めた実験では，距離＝53で平均再現率100%，平均適合率88.71%になる．

3）従来方式に比べ検索精度を大幅に改善することができた．

2）で示した平均適合率88.71%という数字は，6.5で述べたとおり平均再現率100%という厳しい条件下であることを考えると，文字列検索としては十分実用に耐え得るものになっていると言える．

今後の課題は，擬似コードの情報量を出来るだけ少なくすることである．上記1）で提案した擬似コードの長さは144ビットで，1文字の画像ビット数の約1/4になる．この情報量を最小限に押さえるための工夫が必要である．また，本手法をMSゴシックなどの他の文字フォントで書かれた文書や，複数の文字フォントやサイズが混在している文書に適用することも次の課題である．

謝　辞

ペリフェラル特徴の考案者である大阪電気通信大学総合情報学部の梅田三千雄先生には貴重なアドバイスを戴きました．札幌学院大学社会情報学部の佐藤和洋先生には本研究に対する多くの貴重なご意見と助言を戴きました．北大名誉教授の田中一先生には日々暖かな励ましのお言葉を戴きました．また，トランスメディアシステムを開発された北大工学部大学院田中譲研究室の大学院生岡田亮さんにはトランスメディアシステムの利用や実験プログラムを作成する上で大変お世話になりました．記して皆様に感謝申し上げます．

なお，本研究は，第一著者の新國が1999年度の国内留研期間中に行ったものである．本学がこのような機会を与えて下さったことに感謝しています．また，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います．記して感謝致します．

参考文献

秋山照雄，内藤誠一郎，増田功(1984)「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) J 67-D，10，pp.1194-1201

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

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

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，2 H-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

2001年1月22日受付
2001年2月20日受理