TR－A－0070

文書画像検索システム　CHASERS

横澤　一彦
KAZUHIKO　YOKOSAWA

1990. 2. 6

ATR視聴覚機構研究所
〒619-02 京都府相楽郡精華町乾谷　☎07749-5-1411

ATR Auditory and Visual Perception Research Laboratories
Inuidani, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Telephone: +81-7749-5-1411
Facsimile: +81-7749-5-1408
Telex: 5452-516 ATR J

目次

１．まえがき

光ディスクなど大規模記憶媒体の普及によって、そこに蓄積した画像データベースの高速検索が必要になっている。現状の画像検索では、あらかじめ登録したキーワードを用いて行われることが多い。しかしながら、利用形態の高度化に伴い、画像の内容を直接検索する方法が望まれている[97]。ここでは、人間の探索特性を基に、文書画像から特定の文字列画像を抽出し、検索するシステムＣＨＡＳＥＲＳ（CHAracter String image Extraction and Retrival System）について提案する。

２．では、人間の文字探索特性について述べ、３．ではそれを基にした文書画像検索法について述べる。４．では、開発した文書画像検索システムＣＨＡＳＥＲＳの使用法について、解説する。５．では、このシステムの性能評価の為に行ったシミュレーション実験の結果を述べる。

２．人間の探索特性

人間の文字探索特性について心理実験で得られた知見を述べる[2)3)4)]。実験では、図１に示すように、探索目標として、単独文字、単語、非単語のいずれかを提示し、次に探索対象文字列（５０字の文章あるいは無意味文字列）を提示した。被験者は、探索対象文字列中の探索目標の有無について二者択一反応をし、その正答率と反応時間を測定された。反応時間については、図１に示した。実験の結果、平均反応時間に有意差が認められ、対検定を行ったところ、探索目標の有無によらず、単語条件、非単語条件それぞれと単独条件の間で有意差が認められた。更に、無意味文字列で探索目標無の場合に単語条件と非単語条件の間で有意差が認められた。正答率は、単独の時に最も低く、無意味文字列では非単語に比べ単語のときに低くなることが分かった。更に、探索対象文字列中に探索目標と類似した文字を混入すると、その文字数に比例し反応時間が線形に増加した。

すなわち、i)単独文字を探索するより、文字列を探索する方が容易である、ii)非単語に比べた単語の優位性は少ないが、無意味文字列のとき、非単語より正答率が低く、反応時間が速い、iii)大まかな特徴で探索した後、類似文字を識別するような詳細な識別処理が必要であることが分かった。人間が

文書からキーワードを検索する過程において、文字単位の検索処理を行うわけではなく、まずキーワードそのものの大まかな形状特徴を基に検索するものと考えられる。更に、これまでの単語優位効果の分析から、概形処理に用いられる特徴が、外形や複雑さを反映したものであることが示唆されている[5]。実験システムの構成に上記の点を考慮した。

既に、文字形状の観点から、文字列単位の識別の有効性が指摘されている[6]。すなわち、単語をなす前後の文字が確定しているときに、正解文字と２０２０字種の中からランダムに７個、計８個を候補としたならば、９１％の確率で単語は一意に同定される。正解文字と形状類似文字７個、計８個を候補としたならば、８９％の確率で単語は一意に同定される。このように、単語の場合には、形状類似文字に関する詳細な処理をせず大まかな処理で、約９割が一意に同定されることになる。このことは、文字単位処理に比べて、類似文字に対する負担が小さくなると考えられる。従って、これまでにも文字列検索の有効性は指摘されている[7]。この特性を文書画像検索に応用したのが、次に提案する文字列を単位とした直接画像検索法である。

３．文書画像からの文字列検索法

文字列検索のため、従来の文字認識においてもその有効性が確認されている基本的特徴量の中から３つを応用した[8]。すなわち、概形処理で用いる特徴として、外形を反映した１次ペリフェラル特徴[9]、複雑さを反映したストローク密度特徴[10]の２特徴量（図２参照）を、詳細処理で用いる特徴量として、線分の方向分布を反映した外郭方向寄与度特徴[11]（ＰＤＣ特徴、図３参照）を、それぞれ文字列枠に沿って抽出した。詳細処理で用いる特徴についての心理実験からの示唆はないが、従来の文字認識手法の中で高い能力を有する外郭方向寄与度特徴を応用した。しかしながら、予備検討の結果、辞書パターンとの照合において、文字間隔や字体による非線形変形を生ずることが分かった。このような変形を吸収する為ＤＰマッチングを用いて照合した。

このような照合法を用いて、①前処理、②抽出文字列の指定、③概形処理、④詳細処理という検索手順で文字列を抽出する実験システムＣＨＡＳＥＲＳ

を開発した。以下に、それぞれの手順の概略を示す。

①前処理

傾き正規化と行単位の文字列画像抽出を行う。傾き補正における情報に基づき、文書の横書き、縦書きを自動判定する[12]。更に、文書のレイアウト情報をマニュアル入力し、文章、写真、グラフなどの領域分離、文字列の接続関係の自動抽出を行った。従って、複数行にわたるキーワードの抽出が可能である。これらの前処理は、文書入力と同時に行う。

②抽出文字列の指定

検索したい文字列はキーボード入力する。ＪＩＳ第１水準の文字が指定可能であり、その長さは任意に設定できる。更に、タブレットを用いて、手書き入力が可能である。従って、文字だけでなく、記号の検索もできる。次に、このように設定された文字列に対する特徴抽出が行われる。更に、文字列の文字数を基に、候補文字列画像の切出しを行う。左右端（縦書きの場合は、上下端）候補線を、黒画素分布を基に抽出する。左右端の間隔が、文字列画像の高さと指定文字列長から算出される範囲に有れば、候補文字列画像とする。

④概形処理

指定文字列から１次ペリフェラル特徴とストローク密度特徴を抽出する。同様に、候補文字列画像から抽出された２つの特徴量を、それぞれＤＰマッチング法によって照合する。ＤＰマッチング法を用いたのは、変形や字間変動に対処するためである。２つの特徴量それぞれに対する照合結果の閾値処理を順次行う。

⑤詳細処理

指定文字列から外郭方向寄与度特徴を抽出する。同様に、候補文字列画像から抽出された外郭方向寄与度特徴とを、ＤＰマッチング法によって照合し、照合結果の閾値処理を行う。

４．ＣＨＡＳＥＲＳの起動と実行

４．１　ＣＨＡＳＥＲＳの概要

ＣＨＡＳＥＲＳは、２つの部分からなる。すなわち、前処理部と文書検索部である。ハードウェア構成を、図４に示す。図４に示すように、マイクロＶＡＸ、ＶＡＸ８５５０、光ディスク等を使用するので、ＣＨＡＳＥＲＳを起動する為には、それらを立上げておく必要がある。

尚、文書画像データはマスコンプに接続したイメージスキャナで入力し、システム上に転送する。3135x4352のＡ４二値画像とする。

４．２　前処理部

文書画像に対して、傾き補正などの正規化処理や文字列抽出処理などを行う。処理手順は、文書画像ファイルの指定、パラメータの指定、傾き補正、レイアウト構造の指定、文字列抽出、結果の確認の順で進められる。前処理部は、秋山らの手法[12]に基づいている。前処理部の実行方法は以下の通りである。

まず、マイクロＶＡＸにログインし、

```
$  test
$  kanji
```

とキーボード入力し、前処理部を起動する。すると、次のようなメインメニューが表示される。

```
Select file
Pameter
Layout
Batch
Look up result
Exit this menu
```

マウスを使って、このメニューから目的の動作を選択することによって処理を進める。それぞれのメニューの内容は以下の通りである。

(1) Select file

処理すべき文書画像ファイル名を指定する。指定した文書画像ファイル名は、画面左上に表示される。

(2) Pameter

処理を行う際のパラメータの設定、変更を行う。特別の文書画像を扱うとき以外は、変更する必要はない。

(3) Layout

図表、写真部分の指定、段組など文書構造の指定などを行う。この処理は、Batch処理によって、傾き補正が済んでいる文書画像ファイルに対して、行う。具体的には、図５のようにマウスで図表部分を矩形状に指定する。

(4) Batch

傾き補正、文字列抽出などを、多段階に分けてBatch処理する。以下のような処理段階がある。

傾き補正（ＧＰＰ）

傾き補正（ＬＰＰ）

罫線抽出

領域分割線抽出

文字列領域抽出

指定により、これらすべての処理を一括して実行することも、ばらばらに実行することも可能である。

(5) Look up result

Batch処理した各段階の処理結果を表示する。

(6) Exit this menu

前処理部を終了する。

前処理によって得られた文字列画像の例を図６に示す。この前処理部は、単独でも使用可能であり、文書画像処理研究に利用できる。

４．３　文書検索部

前処理された文書画像に対して、指定したキーワード画像の探索処理を行う。処理手順は、文書画像ファイルの指定、パラメータの指定、検索、結果の確認の順で進められる。文書検索システムの実行方法は以下の通りである。

まず、マイクロＶＡＸにログインし、

$　ｔｅｓｔ

$　ｔａｎｇｏ

とキーボード入力し、文書検索部を起動する。すると、次のようなメインメニューが表示される。

Select file
Printing type
Parameter
Searching area
Search
Original image
Batch
Exit

図７に、本システムの表示画面の一例を示す。図中上部は、メニューウィンドウ、図中右は特徴量の分布を表わしている。

マウスを使って、メインメニューから目的の動作を選択することによって処理を進める。それぞれのメニューの内容は以下の通りである。

(1) Select file

処理すべき文書画像ファイル名を指定する。指定した文書画像ファイル名は、画面左下に、１／６４の文書画像がその上に表示される。

(2) Dictionary

三種類の辞書、すなわちPrinting、Hand writing、Tabletを使い分ける。PrintingはＪＩＳフォントを基に作成した活字文書用の辞書、Hand writingは１００人分の手書き文字を基にした手書き文書用の辞書、Tabletはタブレットから入力したパターンを辞書として用いる場合である。

検索文字列は、左上に表示されるSearched Stringウインドウをマウスで選択し、キーボードもしくはタブレット入力する。キーボード入力の場合は、かな漢字変換によって、検索文字列を決定する。

(3) Parameter

処理を行う際のパラメータの設定、変更を行う。特別の文書画像を扱うとき以外は、変更する必要はない。

(4) Searching area

文書検索する範囲を指定する。一行、部分指定、文書画像全体から選択す

る。

(5) Search

指定した文書画像ファイルの指定箇所から、指定した文字列の検索を行う。検索結果は、図７右に示すような散布図の形式で表示できる。

(6) Original image

１／６４画像上で指定した２点で囲まれる矩形位置に相当するオリジナル画像の一部を表示する。図８に、オリジナル画像の１部を示す。

(7) Batch

検索のBatch処理を行う。複数個の文字列の検索を１つのBatch処理で実行できる。Batch処理には、Simulation1とSimulation2の２種類がある。Simulation1では、(5)で述べたSearchと同じ処理をBatch処理する。一方、Simulation2では、検索を行う前に正解位置を入力し、検索時に照合する。従って、結果には正抽出数や誤抽出数などが含まれる。図９に示した例では、右側の円に探索文字列数、正解数、正抽出数が表示されている。

(8) Exit

文書検索部を終了する。

これまで述べてきたのは、マイクロＶＡＸから命令する方法であったが、ＶＡＸのみでBatch処理をすることも可能である。その場合には、あらかじめ検索文字列、パラメータの指定をしておく必要がある。起動方法は、ＶＡＸ８５５０にログインした後、

$　ｓｅｔ　ｄｅｆ　[．ｔｉｓ]

$　ｓｕｂｍｉｔ／ｐａｒａｍ＝文書名　ｂａｔｃｈ１

もしくは

$　ｓｕｂｍｉｔ／ｐａｒａｍ＝文書名　ｂａｔｃｈ２

とキーボード入力する。batch1はSimulation1に、batch2はSimulation2に対応する。

５．評価実験

4352x3136画素の90文書画像から、１画像当り１０種類のキーワードを選び検索能力を検討した。文書は、研究論文や出版物をＡ４用紙に拡大複写し、

それをイメージスキャナで１６本／ミリの２値入力した。従って、特に画像品質が高い訳ではなく、特別な雑音除去も行なわなかった。90文書画像のうち、68画像が横書き、22画像が縦書きである。１画像当り１０種類のキーワードを選択したが、半数が２文字単語であり、残りが３文字から７文字の単語である。３文字以上単語のうち１種類は片仮名もしくは平仮名のみからなる単語とした。更に、２文字単語と３文字以上単語のそれぞれ１種類は２行にまたがるキーワードとした。照合における閾値処理などのパラメータは、予備実験によって決定した。

評価実験の最終結果を表１に示す。ここで行なった評価実験は、検索した候補文字列画像延べ1,758,236文字列画像の中から、正解である1,348文字列画像を選択する課題、すなわち全体から0.077%の正しい情報を抽出する課題である。各キーワードは文書中に複数存在する場合があり、９００以上の文字列画像が正解となる。延べ１７６万文字列画像から、概形処理において15,150文字列画像が選択され（約1/116に候補削減）、詳細処理ではこの文字列画像についてのみ照合した。従って、概形処理を用いず詳細処理だけを行なう場合に比べ大幅に検索時間を短縮することができた。概形処理で選択された画像には正解の９７．４％が含まれている。未検索の２．６％のうち、約半数が漢字を含まない片仮名単語あるいは平仮名単語であった。９００種類のキーワードから漢字を含まない片仮名単語と平仮名単語９０種類を除けば、概形処理で正解の９８．４％を含んでいたことになる。このように、概形処理は候補文字列画像の削減に非常に有効であり、片仮名単語や平仮名単語の抽出に比べ、漢字単語の抽出が容易であることが確認された。ある活字文書画像から抽出した"周波数"、"周"、"波"に対する候補文字列の概形処理における特徴距離分布を図１０に示す。特徴次元数は、文字列の長さによらず同次元の場合である。図１０から明らかなように、"周"に対して"原"、"波"に対して"被"などの類似文字が存在し、１文字では正しい検索が困難でも、文字列"周波数"としての検索ならば、正解文字列とそれ以外の文字列の特徴分布の分離性が高く、２特徴の併用効果もあることが分かった。

詳細処理まで行なった結果では、表１に示す通り、1,491文字列画像が選択され、その中に正解の９２．９％が含まれていた。図１１に検索もれと誤

検索の例を上げる。詳細処理でもれた正解のうち、６割以上が３文字以上の漢字単語であり、文字数に増えるにしたがって正しい抽出が困難になる傾向が見られた。

また、誤検索の中には、正解文字列とほぼ領域を共有する次のような２種類が含まれている。すなわち、正解文字列＋（ごみ、括弧、平仮名の一部など）と正解文字列－（先頭文字か最後の文字の一部）である。これらは誤検索の１１．３％を占めるが、このすべてが正解文字列と共に抽出されている。それ以外にも、正解文字列と一部分が同じ文字列画像が誤検索される（例えば、”全体”や”類似文字”を検索させたときに正解文字列と共に”合体”や”類似漢字”が誤検索される）場合があり、これらは誤検索の６４．４％を占めた。

文字列単位の検索をすることで、類似文字に対する処理は軽減されるはずであるが、ここで用いた詳細処理は実用的にはまだ充分でないことを示している。従って、詳細処理で用いる特徴や照合法について、今後更に心理学的実験も含め検討する必要があろう。例えば、単語優位効果の分析では詳細処理において文字単位のシリアルな処理が示唆されている。概形処理後に残された、正解文字列と形状が類似した候補文字列中の特定文字に重み付けした照合が必要であると考えられる。

表１

		枚数	検索単語	正検索	未検索	誤検索	候補単語
横書き	活字	42	678	640	38	103	976970
	ﾄﾞｯﾄﾌﾟﾘﾝﾀ	26	354	320	34	80	558969
縦書き	活字	22	316	292	24	56	222297
合計		90	1348	1252	96	239	1758236

６．むすび

文字と文字列の探索過程を調べた心理実験から、文脈中の文字認識処理は、文字単独の認識処理とは異なり、並列処理によって高速で効率的処理が行われていることが明らかになっている。これは、概形処理と詳細処理の２段階処理モデルによって説明できる。

このような結果を基に、文書画像から任意の文字列画像を検索する実験システムＣＨＡＳＥＲＳを構築した。概形処理としては複雑さと外形を反映した特徴を、詳細処理としては線分の方向分布を反映した特徴を抽出し、文字変形と文字間隔変動に対処する為、ＤＰマッチング法を照合に用いた。この実験システムＣＨＡＳＥＲＳによって、９０文書画像から９００種類の単語を検索する評価実験を行い、その有効性を確認した。特に、概形処理において、１文字では正しい検索が困難でも、文字列の検索ならば、特徴分布の分離性が高いことなどが分かった。提案した手法は画像の直接検索法であるので、記号などの検索や文書筆記者自身の手書き入力による検索も可能である。

謝　辞

この報告書は、ＡＴＲにおいて著者が中心になって進めてきた『文字認知機構の研究』から、文書処理システムについて解説したものである。このような研究の機会を与えていただいたＡＴＲ視聴覚機構研究所淀川英司社長に感謝致します。また、日頃御討論頂く視聴覚機構研究所の諸氏にも感謝致します。

参考文献

1)坂内正夫:"画像検索技術",信学誌,71,9,911-914,1988

2)横澤一彦:"日本語の視覚的処理単位－単語認識過程における諸現象－",ATRテクニカルレポート,TR-A-0066,1990

3)横澤一彦:"文書画像中の文字列検索に関する検討",1989春季信全大,D-473

4)横澤一彦:"文字探索課題における単語優位性",1989基礎心理学会大会

5)横澤,森,梅田:"単語認識過程における文脈効果－単語優位効果と単語文字現象－",TV学会視覚情報研究会,VVI'87-22,1987

6)梅田三千雄:"マルチフォント印刷漢字の分類",信学論D,62-D,2,133-140,1979

7)大田,鈴木,池田:"手書き日本語文認識における文字列利用の一方式",信学論,J68-D,3,330-336,1985

8)橋本新一郎編著:"文字認識概論",電気通信協会,1982

9)梅田三千雄:"単語辞書を用いた文字認識における文字の確定能力",信学論,J72-D-II,1,22-31,1989

10)内藤,淀川:"手書き漢字のストローク密度関数による大分類",信学技報,PRL79-3,1979

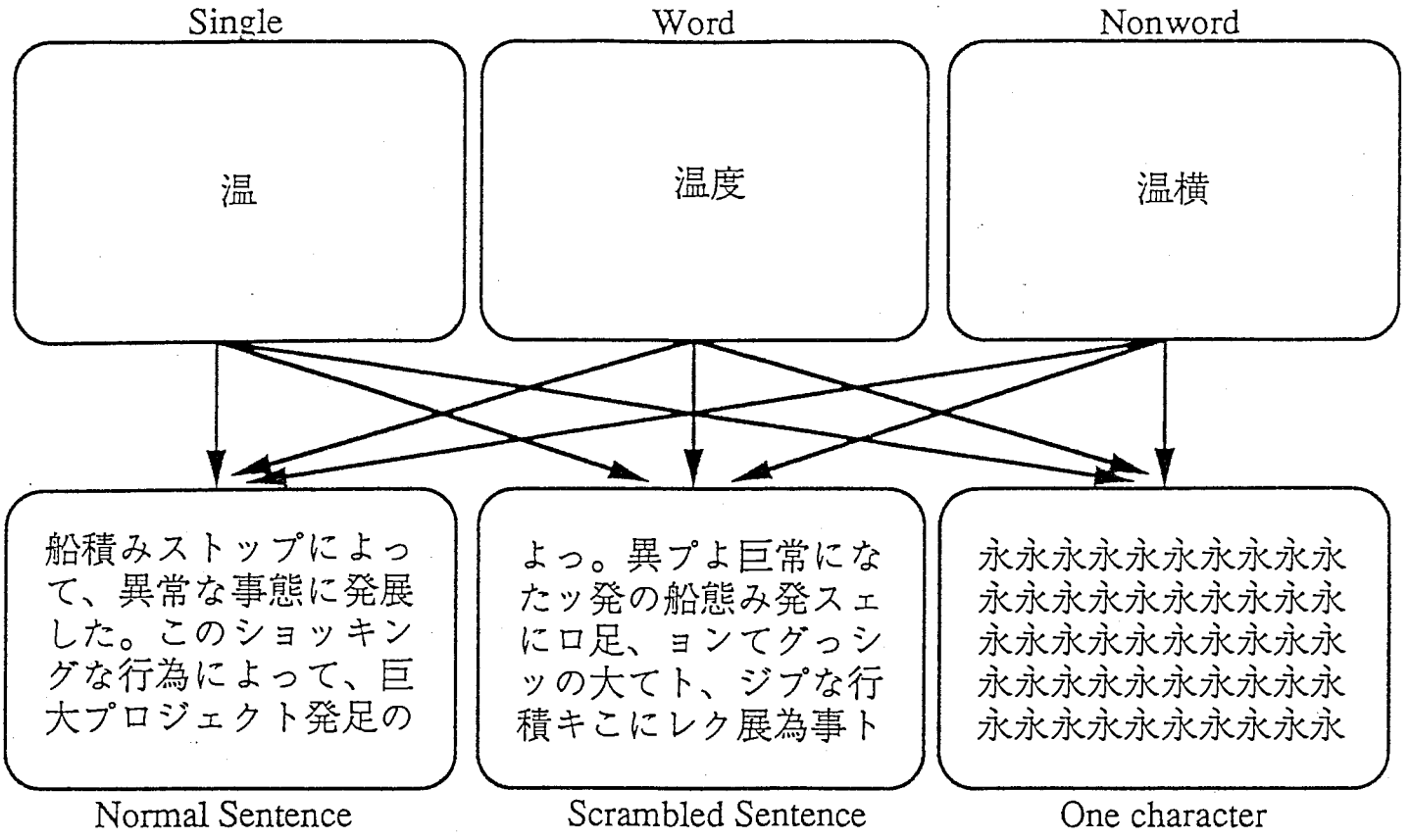
11)萩田,内藤,増田:"外郭方向寄与度特徴による手書き漢字の識別",信学論,J66-D,10,1185-1192,1983

12)秋山,増田:"書式指定情報によらない紙面構成要素抽出法",信学論,J66-D,1,111-118,1983

13)横澤一彦:"人間の文字探索特性に基づいた文書画像中の文字列検索",1989秋季信全大,D-182

14)K.Yokosawa:"Human-based character string image retrieval from textual images",Proc.of IEEE International Conference on System Man & Cybernetics,1068-1069,1989

15)横澤一彦:"人間の文字探索特性とそれに基づく文書画像検索",信学論D-II採録予定,1989



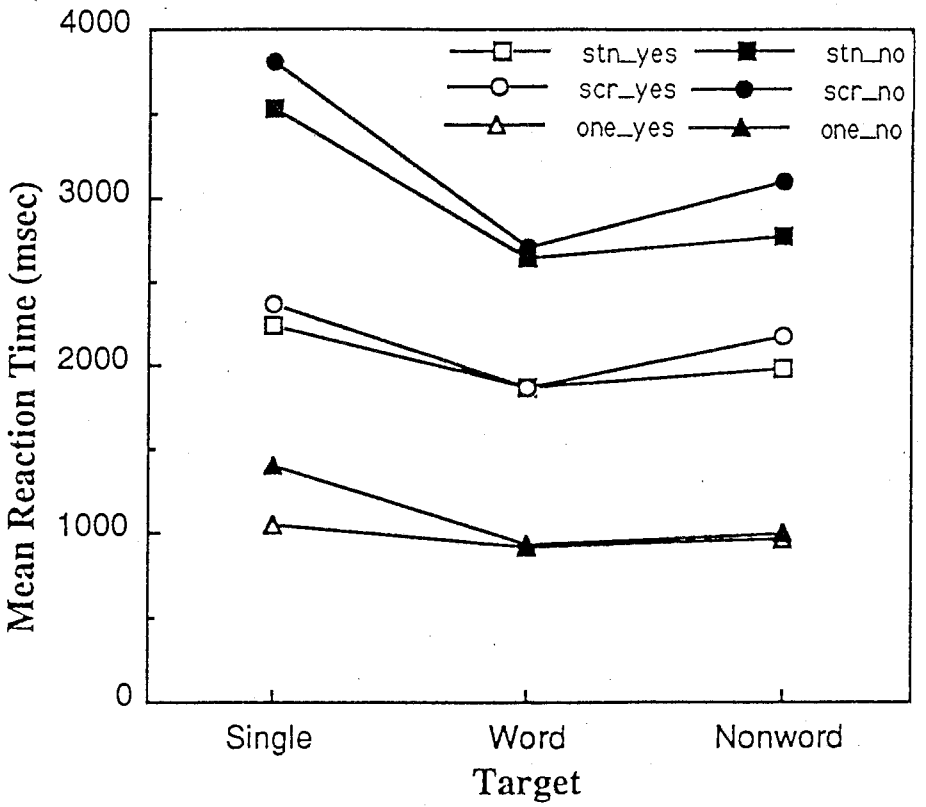


図 1　文字列探索に関する心理実験

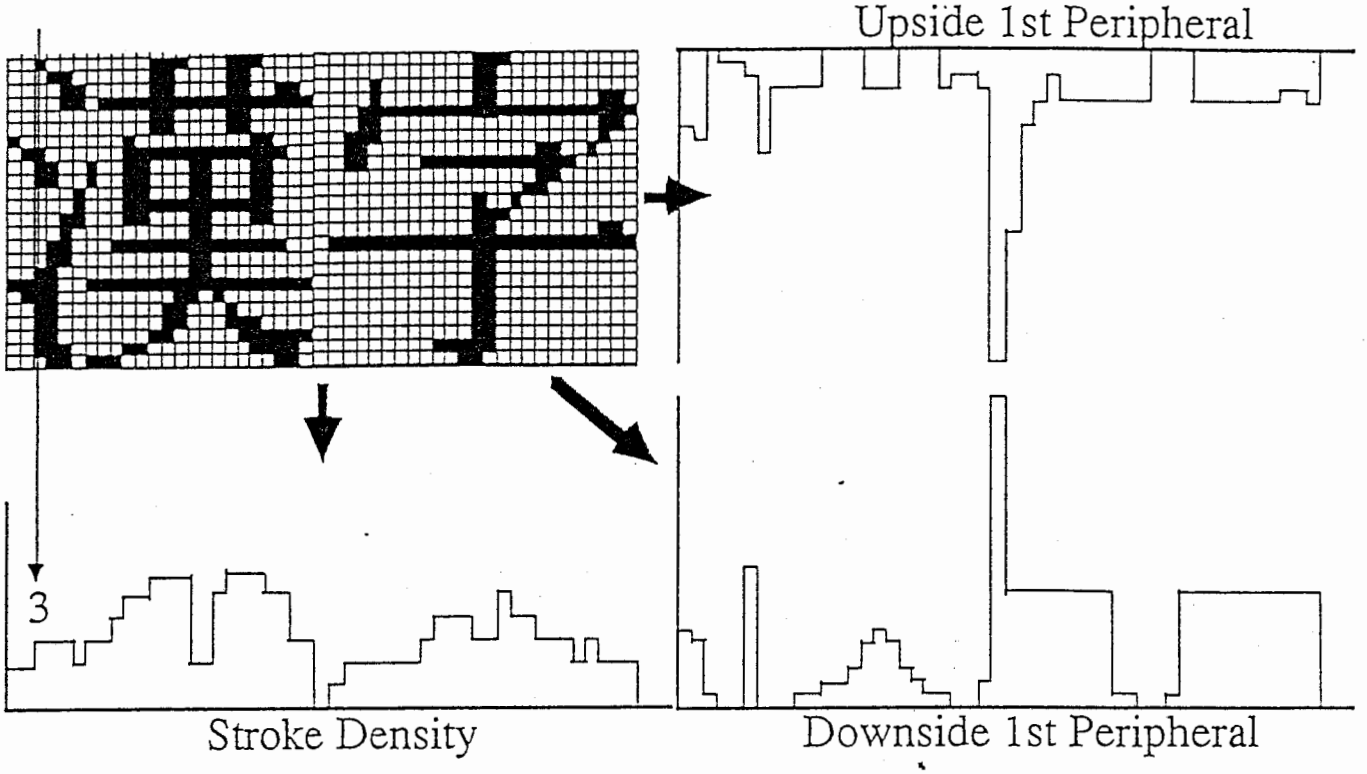


図２　概形処理で用いた特徴量

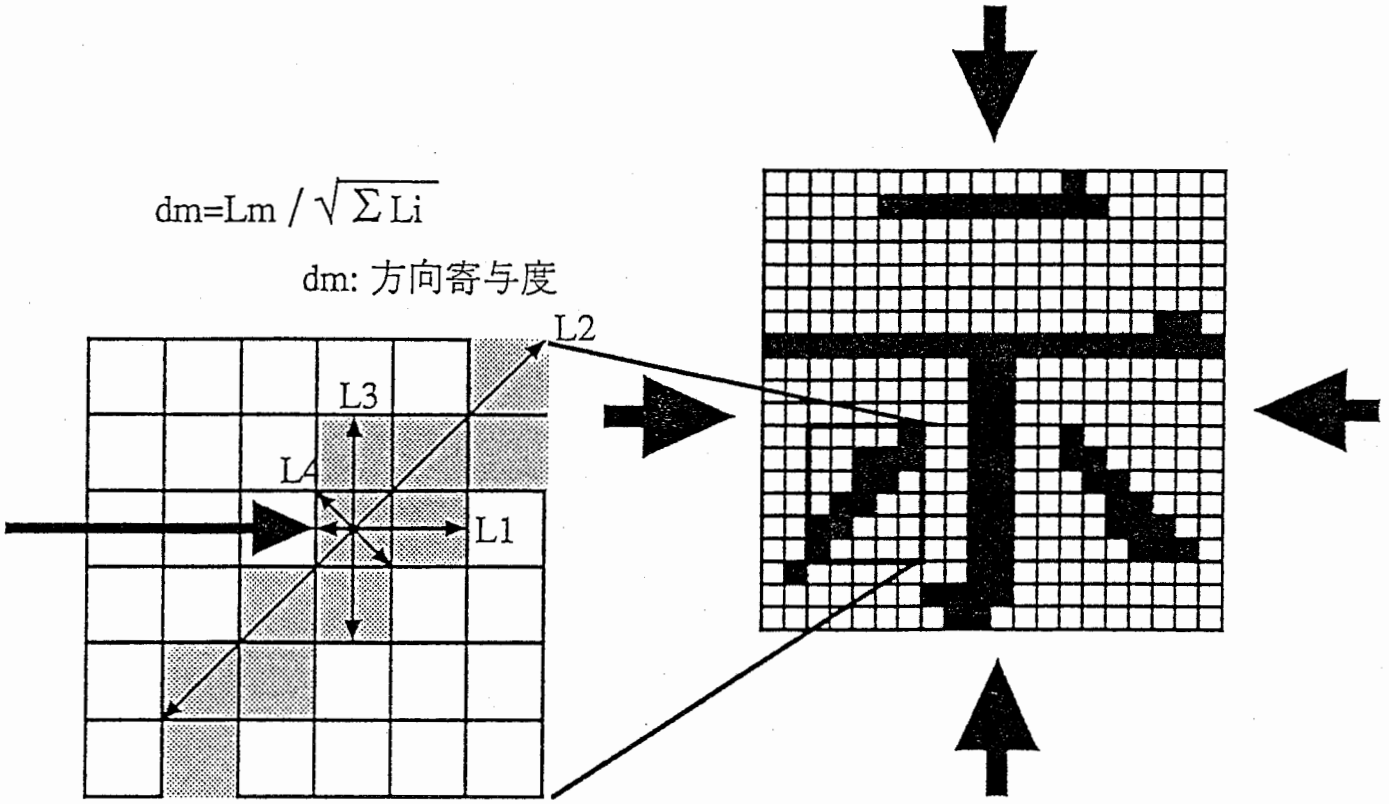


図３　詳細処理で用いた外郭方向寄与度特徴

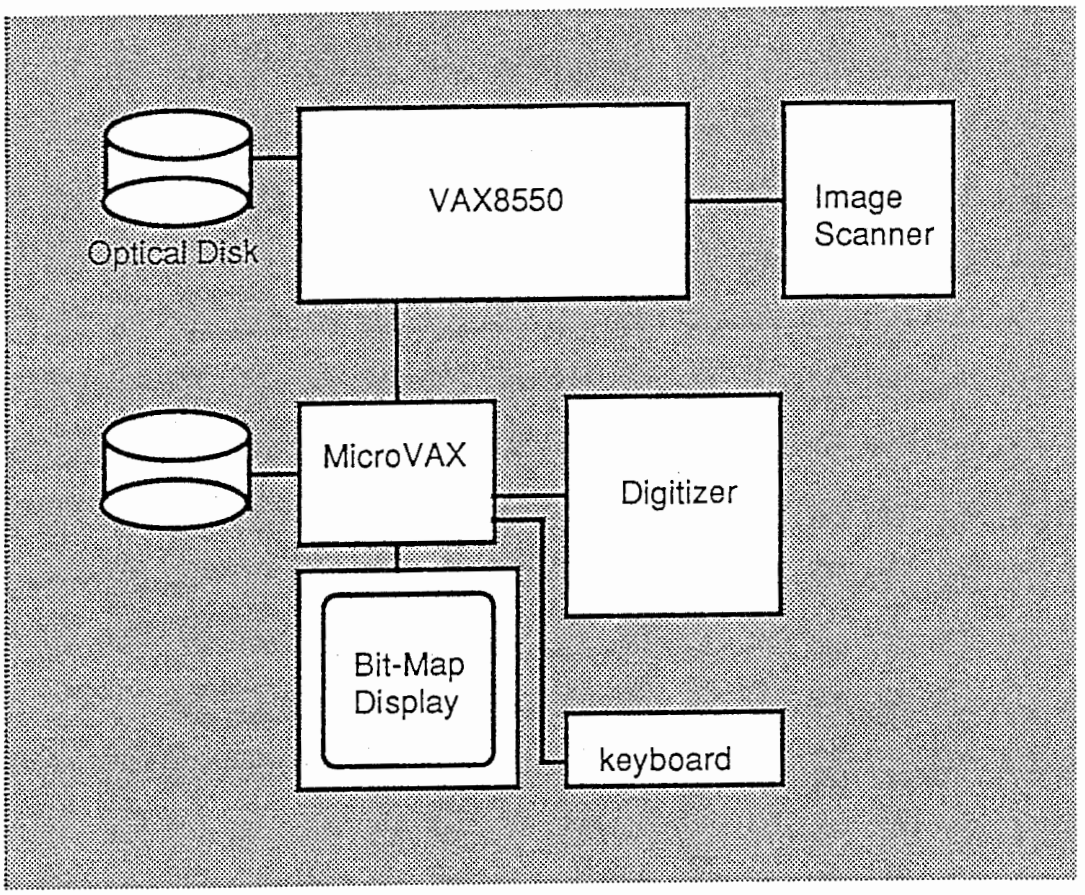


図４　文書検索システムのハードウェア構成

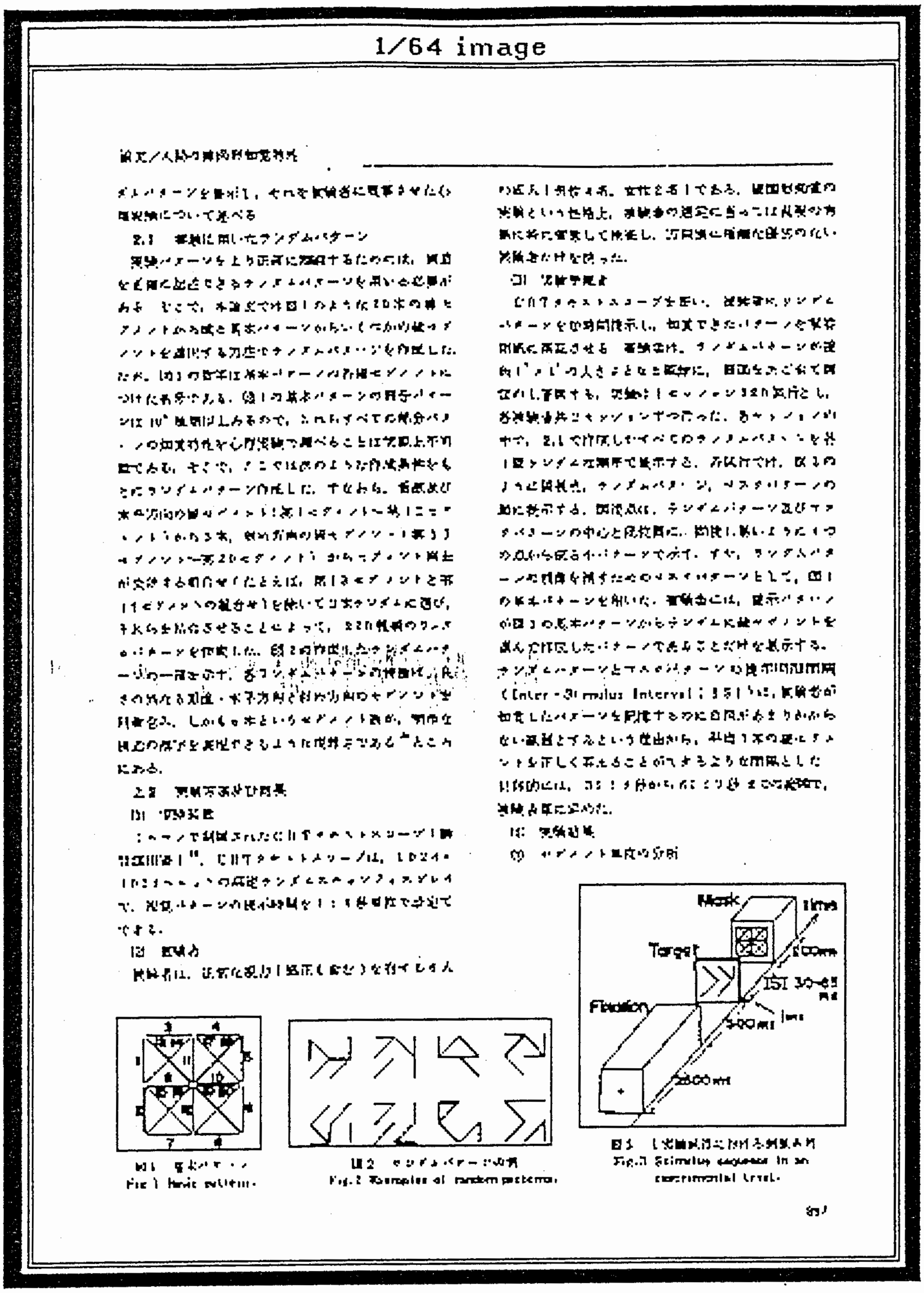


図５　前処理における図表の指定

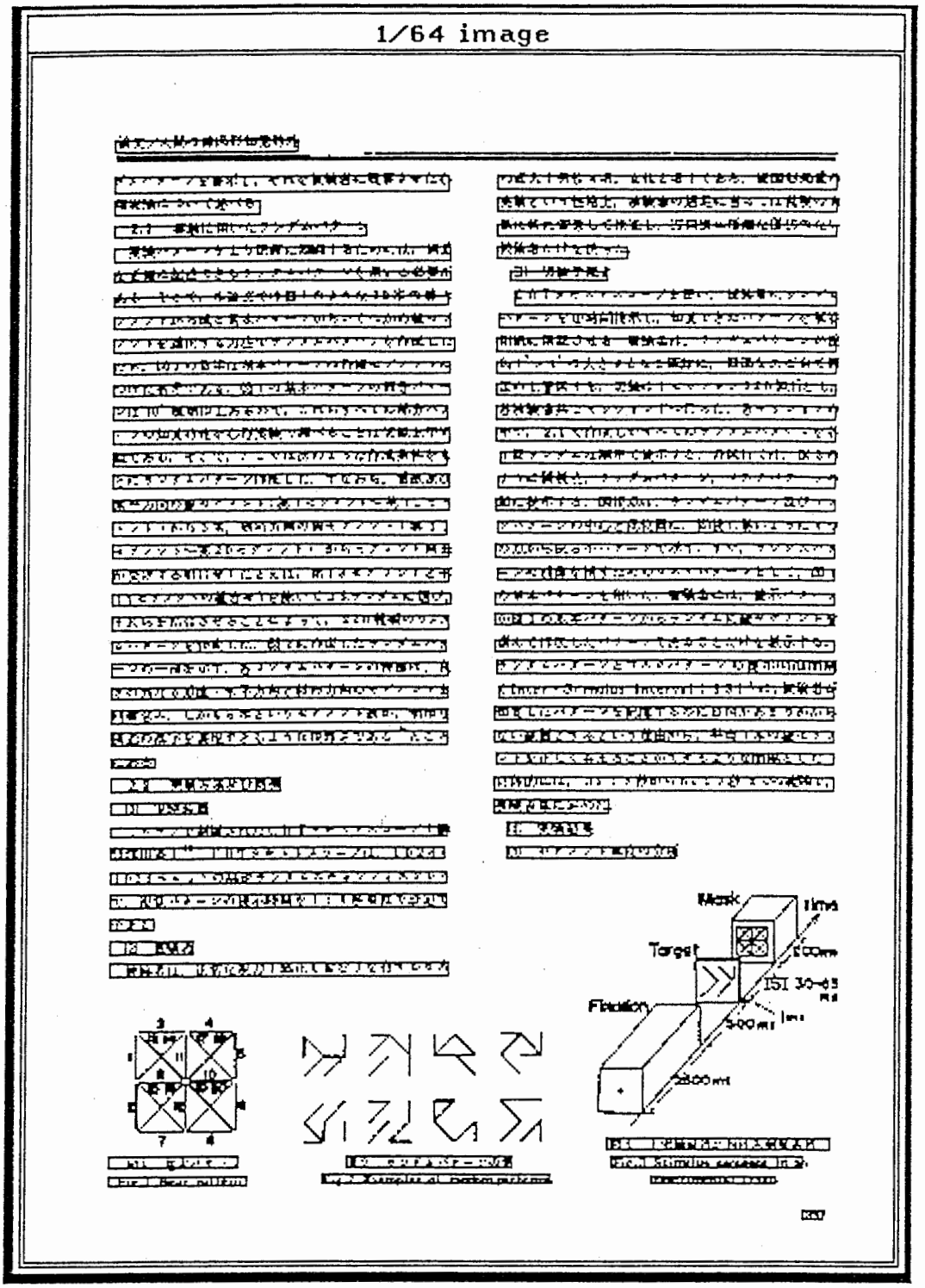


図６　前処理で抽出された文字列画像

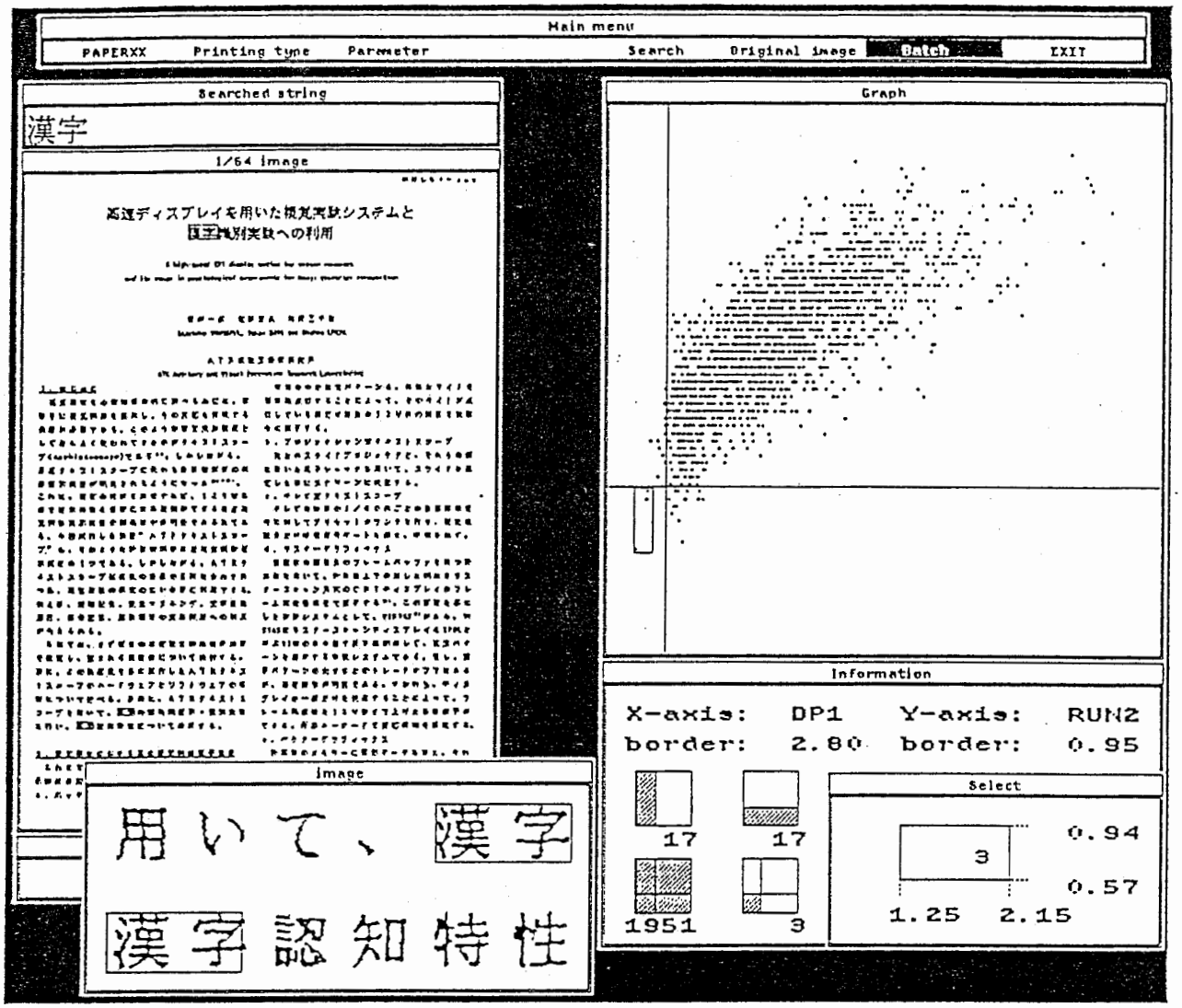


図７　検索処理後の表示画面例

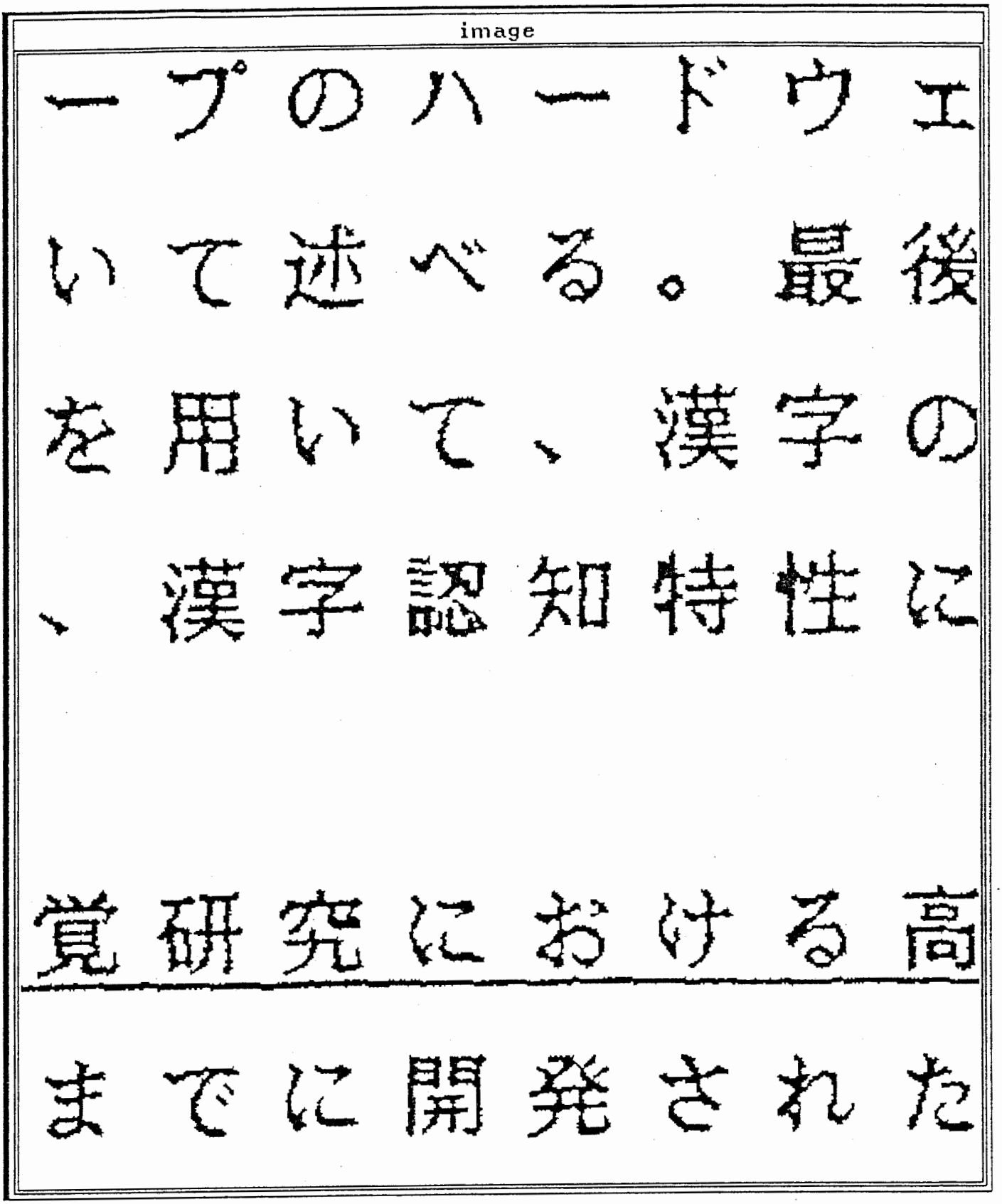


図８　オリジナル画像の一部

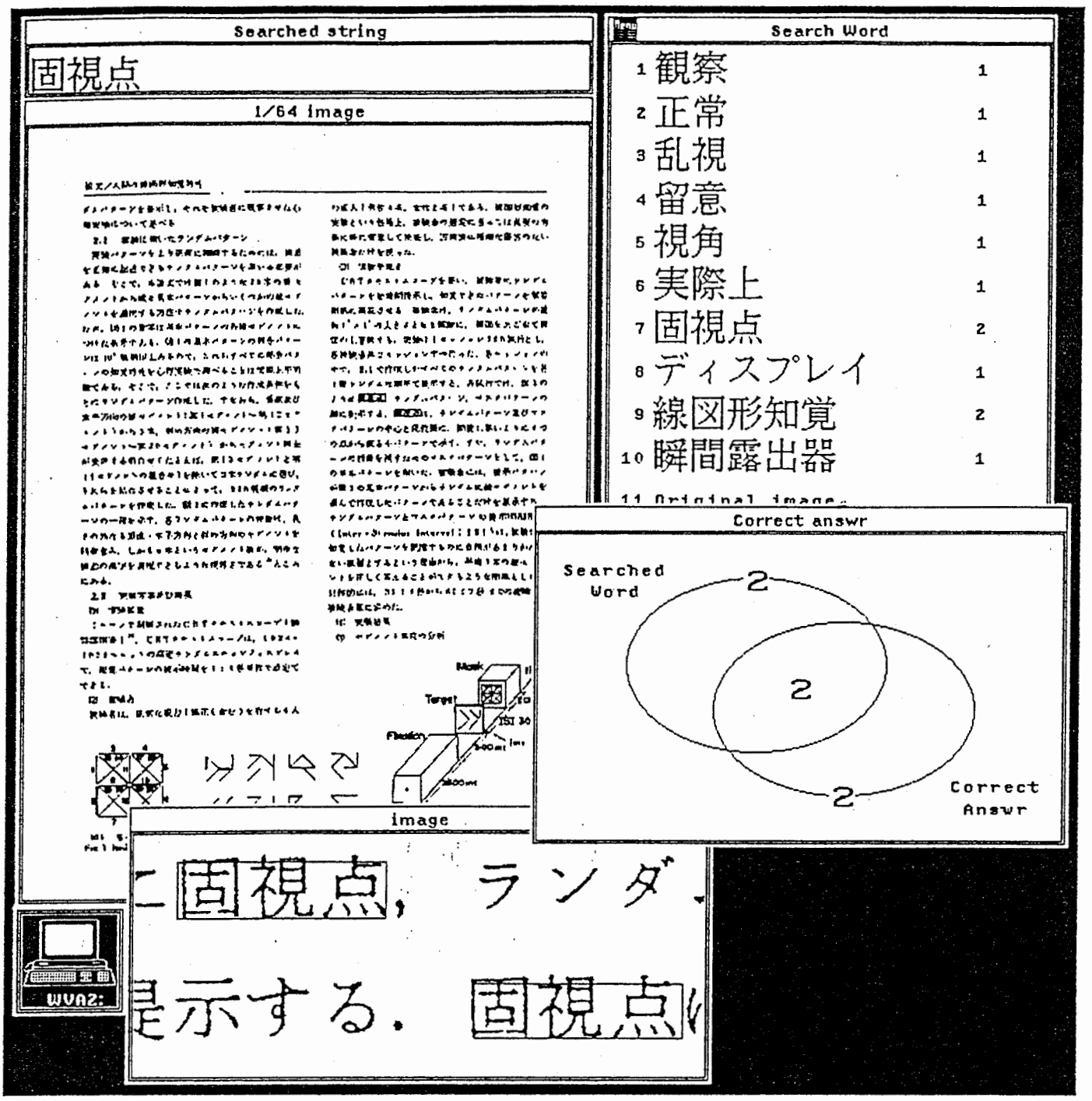


図９　正解との照合を含む検索処理後の表示画面例

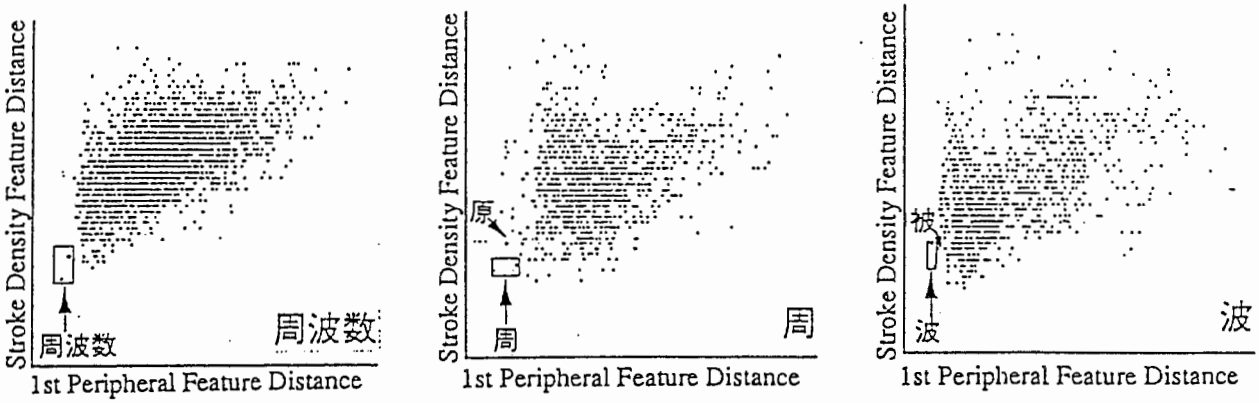


図１０　概形処理における特徴分布

検索もれ

縦横　黒画素

誤検索

に反比例し，X_3 に比例する

分から全体
(top-down
に基づいた
合体する(，
音符と意符

デメリット

「書道」

円滑

統一的に

情報処理過程

図１１　検索もれと誤検索の例