97-01067

部分画像クラスタリングに基づく文書画像の構造化と検索＼＼

黄瀬　浩一　大阪府立大学大学院工学研究科電気・情報系専攻助教授

1　研究の目的

コンピュータやインターネットの普及に伴って、大量の文書が電子的に作成されるようになってきた。ところが、我々はしばしば、電子文書を紙に印刷して閲覧する。これは、両者の特性の違い、すなわち、電子文書は加工や検索などに優れているが、印刷文書は閲覧、携帯などに優れていることによると考えられる。その結果、現在でも印刷文書は大量に作成され続けている。また、図書館などには対応する電子文書が存在しない印刷文書も大量に保管されており、価値のある情報が記録されていることも少なくない。このような状況を考えると、印刷文書の代替として電子文書を考えるよりは、両者の共存が望ましいといえる。

印刷文書を扱う上での大きな問題は、保管に大きなスペースが必要であること、ならびに、記録された情報を効率的かつ柔軟に検索できないことであろう。これらの問題点は印刷文書の電子化により解決される。

直接的な電子化の方法は、OCRを用いて電子文書へ変換し、電子文書データベースを作成することであろう。この方法を用いると既存の電子文書検索が可能となるという利点が得られるが、入力時のコスト(OCRの誤りの訂正)が大きな問題となる。この問題点を回避しつつ電子化する方法として、印刷文書を文書画像という形でデータベース化すること(文書画像データベースの構築)が考えられる。ところが、この方法には、データベースに格納されているものがあくまでも画像であり、電子文書と同様の検索を施すことができないという問題点がある。

そこで、本研究では、文書画像データベースの持つ上記の問題点を解決することにより、電子文書データベースと同様に柔軟な検索を実現することを目指す。具体的には、次の2点について検討した。

1　文書画像のレイアウト解析［1, 2, 3, 4］＼＼

　電子文書は1次元的な記号の並びであるのに対して、文書画像は2次元的な広がりを持つものである。 レイアウト解析とは、文書画像を、ブロック（テキストブロック、図表など)、文字列などの単位へと切り分ける処理であり、その成否が後の処理に大きな影響を及ぼす。本研究では、従来法では困難であった、複雑なレイアウトを持つ文書画像の解析法を検討した。

2　文書画像データベースの検索［5, 6］＼＼

　電子文書の検索と同様の利便性を得るためには、文書画像に対するキーワード検索を実現する必要がある。これを実現する上での最も大きな問題は、ユーザから与えれるキーワードが記号であるのに対して、画像が信号でしかない点である。本研究では、記号を信号(特徴ベクトル)に変換することにより、このギャップを埋めて検索を行う手法を検討した。また、このとき、処理性能と速度を向上させるため、部分画像のクラスタリングという考えを導入した。

2　文書画像のレイアウト解析

2.1　複雑なレイアウト

従来のレイアウト解析では、主に、テキスト、図表などの文書の構成要素が、矩形の領域を持つことを前提としてきた。このようなレイアウトは矩形レイアウトと呼ばれるものである。対象文書のレイアウトを矩形レイアウトに制限することにより、従来法では、処理を高速かつ信頼性の高いものにすることが可能となっている。ところが、近年、矩形レイアウトの枠に収まらないようなレイアウトを持つ文書が増えている。一例を図1(a)に示す。この例では、左上部に他と異なる傾きを持つ文字列が円状に配置されている。

このようなレイアウトを持つ文書も文書画像処理において同様に扱うためには、入力文書のレイアウトに関する仮定を排除した柔軟なレイアウト解析法が望まれる。

2.2　レイアウト解析の指針

レイアウトに関する仮定を排除するため、本手法では、文書画像の連結成分を処理単位とする。連結成分とは、縦横斜

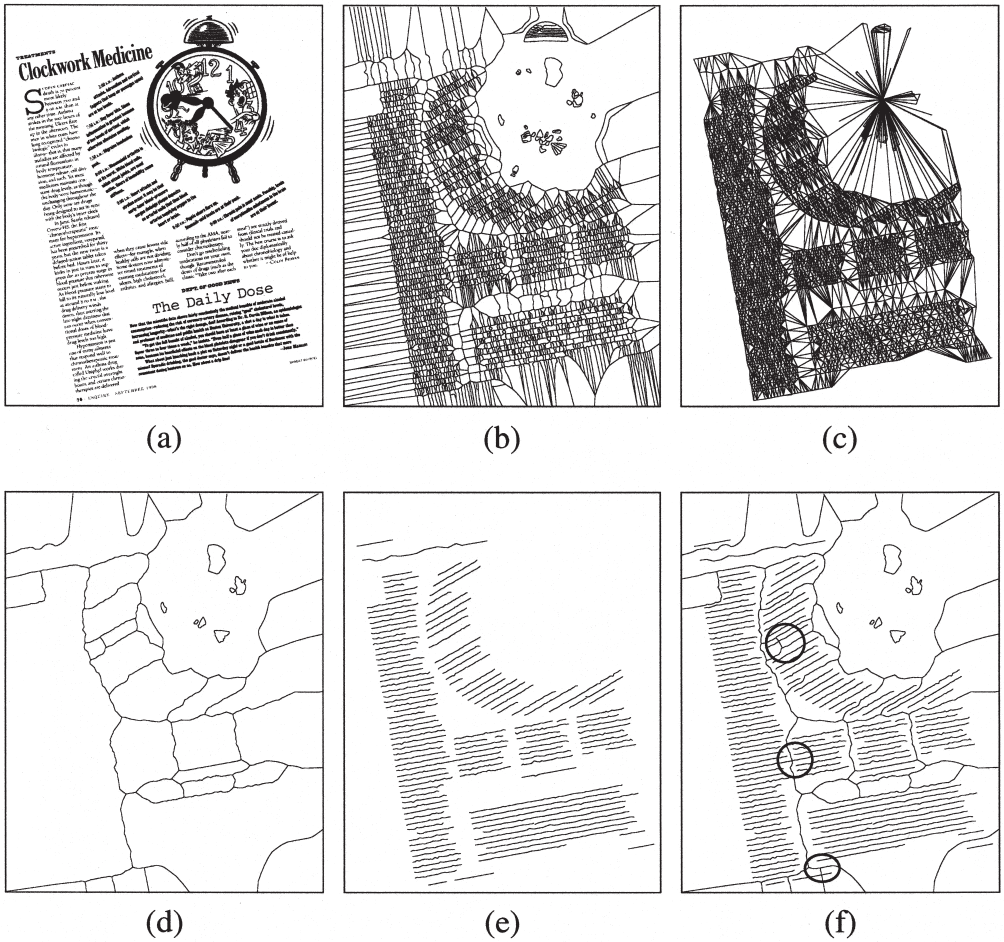

図1　処理プロセス。(a)入力画像、(b)一般図形ボロノイ図、(c)隣接グラフ、(d)抽出されたブロック、(e)抽出された文字列、(f)(d)と(e)を重ねたもの。

めに隣接する黒画素をまとめたものである。そして、連結成分をグループ化することにより、ブロックや文字列を構成する。

グループ化に際しては、次のような単純な方針を用いた。

1　ブロックをグループ化する指針

互いに近接する連結成分は同じブロックに属する可能性が高い。

同程度の大きさの連結成分は同じブロックに属する可能性が高い。

2　文字列をグループ化する指針

互いに近接する連結成分は同じ文字列に属する可能性が高い。

直線状に並ぶ連結成分は同じ文字列に属する可能性が高い。

同程度の大きさの連結成分は同じ文字列に属する可能性が高い。

2.3　ボロノイ図の利用

一般に、文書画像に含まれる連結成分の数は非常に多い。したがって、上記のような方針を実現するには、注意が必要となる。すなわち、連結成分のすべての組み合わせに対して近接性や直線性、大きさの類似性を計測することは処理時間から現実的ではない。

この問題を解決するため、本手法では、ボロノイ図を利用する。

通常、ボロノイ図は、与えられた点(母点)に基づく領域分割を表すものである。分割された領域内の点は、その領域に属する母点に最も近いという性質を持つ。計算量は母点の数をnとすると$O(n\ddot{A}log\ n)$であることが知られている。

本手法では、連結成分に対するボロノイ図を作成する。連結成分は大きさを持つものであるため、上記の通常のボロノイ図(点ボロノイ図)を適用することができず、一般の図形に対するボロノイ図(一般図形ボロノイ図)が必要となる。本手法では、点ボロノイ図を用いて近似的に構成される一般図形ボロノイ図を用いる。

例を図2に示す。図2(a)は元の文書画像、同図(b)は連結成分の境界点から構成された点ボロノイ図である。同図(c)は(b)の点ボロノイ図のうち、同じ連結成分に属する母点から得られた辺を除去したものである。このような処理により、一般図形に対するボロノイ図を近似的に構成できる。

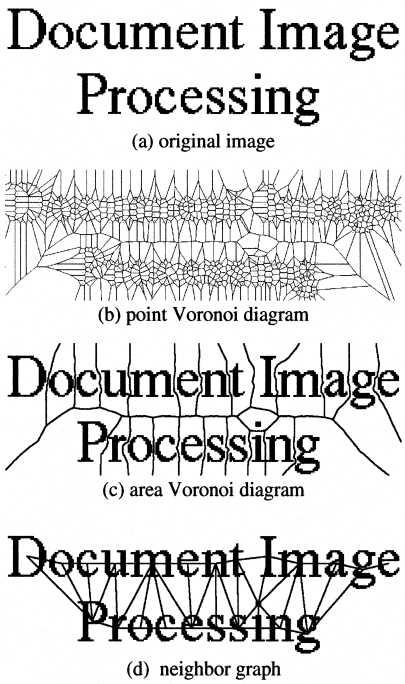

図2　一般図形ボロノイ図と隣接グラフ

連結成分の一般図形ボロノイ図については、次のことがいえる。連結成分のうち、領域の境界(ボロノイ図の辺)が隣り合うものは、隣接する連結成分と捉えることができる。2.2で述べた近接性や大きさの類似性を計測するとき、互いに隣接する連結成分に対象を絞れば、計算量を削減することができる。

図2(d)は、連結成分間の隣接関係を明示的にグラフで表したものである。これを、隣接グラフと呼ぶ。文字列抽出の指針で述べた「直線状に並ぶ」という基準については、隣接グラフの辺の並びが直線的かどうかにより評価することができる。

2.4　レイアウト解析法の概要

以上に述べた方針に沿って、本手法は実現されている。ここでは、図1の処理例に基づいて概要を説明する[1]。

図1(a)は入力文書画像であり、10 ° の傾きを持つ。解像度は300dpiである。この画像には、5.4×10^3個の連結成分が含まれている。黒画素の個数は1.08×10^6である。図1(b)は一般図形ボロノイ図である。この図は、連結成分の境界点6.0×10^4個から作成された。辺の総数は、8.3×10^3である。図1(c)は対応する隣接グラフである。辺の数は図1(b)と同じである。図1(d), (e)は、それぞれ抽出されたブロック、文字列を表す。ブロック、文字列ともに、一部、誤抽出や未抽出が見受けられるが、おおむね正しい結果が得られている。図1(f)はこれら2つを重ねたものである。現在は、ブロック抽出と文字列抽出は独立に適用されているため、一部に矛盾する結果(図1(f)の太線で示す部分；文字列がブロックを跨いでいる)が得られている。これを解決するためには、矛盾を認定し、回避する処理を追加する必要がある。

最後に処理時間について述べる。Pentium II 450MHzのPCを用いて計測したところ、ブロック抽出に3.4秒、文字列抽出に5.2秒を要した。一般図形ボロノイ図や隣接グラフを得るための計算は、1.4秒であった。以上から、本手法は、複雑なレイアウトの文書を高速に処理可能であることがわかる。

[1]処理の詳細については、ブロック抽出は［2］、文字列抽出は［3, 4］を見よ。

3　文書画像データベースの検索

次に、本研究のもう1つの目標である、文書画像データベースの検索について述べる。

文書画像データベースとは、文書を画像として格納したデータベースである。ユーザは文書画像データベースから所望の文書を検索する際、自身の要求を検索質問として与える。検索質問としては様々なものが考えられるが、ここでは、キーワードのリストを与える場合を考える。検索処理では、与えられた検索質問に適合する文書画像を選択し、ユーザに提示する。

以下、従来法とその問題点を整理したあと、提案手法について述べる。

3.1　従来法と問題点

従来の文書画像データベースの持つ問題点は、大きく次の2つに分類できる。

1　インデックス付けの問題点

インデックスとは、個々の文書画像を特徴付ける指標である。検索処理では、検索質問とインデックスを照合することにより、文書画像を取捨選択する。

最も単純な方法は人手によるインデックス付けである。つまり、対象文書画像を特徴付けるキーワードを人手で付与するものである。ところが、この方法はコスト面から現実的ではない。

この問題を解決するため、これまでに様々な自動インデックス付けの試みがなされている。通常は文書画像から文字を切り出したあと、個々の文字画像に処理を施すことによりインデックスを付与する。 ところが、レイアウト解析、文字切り出しは必ずしも成功するとは限らないため、問題が生じる。

2　検索の問題点

ほぼすべての従来法では、キーワードが個々の文書画像中に存在するかどうかを吟味し、結果を画像単位で出力するものである。このような処理では、文書画像が単一のトピックを表すものであることが前提となっている。

ところが、文書には、いわゆるマルチトピックのものがある。例えば、新聞のページが画像として保存されている場合、ページ単位の提示は、かならずしも適当であるとはいえない。これは、ページ単位で出力されても、ユーザはその中から自身の必要とする部分を探さなければならないためである。

同様の問題は、電子文書の検索でも生じており、その解決法としてパッセージ検索(passage retrieval)が 注目されている［7, 8］。パッセージ検索とは、取捨選択の単位を文書ではなく、その部分(パッセージ)とするものである。これにより、ユーザの利便性や検索精度の向上が期待されている。

したがって、文書画像の検索においても、同様のことを考える必要がある。

3.2　処理の概要

以上の問題点を考慮し、本研究では、文字切り出しを必要としないインデックス付与、文書画像に対するパッセージ検索という2つの特徴を持つ検索法を考案した。これらは具体的には以下のような特徴である。

1　文字切り出しを必要としないインデックス付与

文書画像から容易に抽出できる連結成分を単位としてインデックスを付与する。インデックスとしては、記号ではなく、連結成分を解析して得られる特徴ベクトルを付与する。

2　文書画像に対するパッセージ検索

本手法では、検索質問として与えられたキーワードの密集して存在する部分が、求める部分であるという考え方に基づいてパッセージ検索を実現する。

キーワードの密集度合は、出現密度という値により表される。出現密度を閾値処理することにより、文書画像のなかから該当部分を特定する。

処理の流れを図3に示す。処理は大きく、文書画像データベースの構築と検索に分類される。文書画像データベースの構築では、(1)まずスキャナにより印刷文書を文書画像に変換し、(2)文書画像からインデックスを生成する。
また、検索では、(3)検索質問を検索キーと呼ぶ特徴ベクトルに変換し、(4)それをインデックスと照合することにより、文書画像から適合部分を得る。

以下、各々の処理について述べる。

3.2.1　文書画像データベースの構築

手順に沿って概要を示す。

(1) 部分画像の抽出

文書画像から連結成分を抽出し、面積に基づいて文字の構成要素としては大きすぎるもの、小さすぎるものを除去する。残った連結成分に対して外接矩形を考え、それらが重複するものを統合する。この結果得られるものを部分画像と呼ぶ。

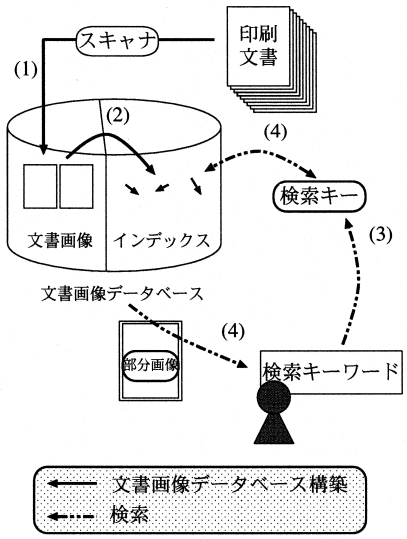

図3　文書画像データベースの概要

⑵特徴ベクトルの抽出

個々の部分画像からメッシュ特徴を抽出し、KL展開による次元圧縮を施す。その結果、部分画像を表す32次元の特徴ベクトルが得られる。

⑶インデックス付け

特徴ベクトルに基づいて部分画像をクラスタリングすることにより、類似の画像を1つのクラスタにまとめる。また、各クラスタについてクラスタ重心のベクトルを求め、これをインデックスとする。

文書画像データベースには、インデックスの特徴ベクトルと、クラスタに属する個々の部分画像の位置情報をペアにして保存する。

3.2.2　検索

⑴特徴ベクトルへの変換

記号として入力されたキーワードから個々の文字を得る。そして、文字に対して特徴ベクトルの辞書を参照することにより、インデックスと同様の特徴ベクトルに変換する。一般には、1つの文字から複数の特徴ベクトルが得られる。以下では、キーワードから得られた特徴ベクトルを検索キーと呼ぶ。

⑵検索キー出現密度の計算

まず、検索キーとインデックスのユークリッド距離を求める。距離がある閾値を下回るとき、両者は類似していると考える。

次に、そのようなインデックスに対して、距離に応じた高さを持つ重み関数を用いて、画像空間に投票する。重み関数としては、図4に示す2次元ハニング窓関数を用いる。投票は、重み関数の中心をクラスタに属する部分画像の中心にあわせ、重み関数の値を画素に加えることにより行う。

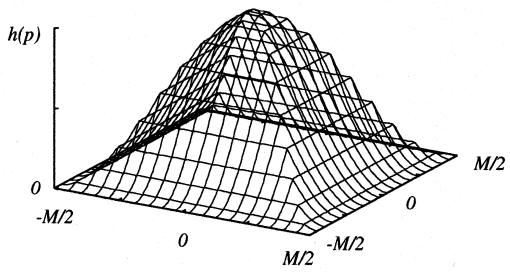

図4　重み関数（2次元ハニング窓関数）

⑶キーワード出現密度の計算

一般に、キーワードは複数の検索キーから構成されている。そこで、次に、各検索キーの出現密度に基づき、キーワー

ドの出現密度を求める。このとき、キーワードは各検索キーがすべて存在するときに、はじめて存在するとみなせることから、ANDに相当する演算をするのが妥当であると考えられる。そこで本手法では、検索キーの出現密度を最大値で正規化したのち、各位置ですべての検索キーの出現密度の最小値を取ることにより、キーワード出現密度とする。

⑷検索質問出現密度の計算

検索質問は、一般に複数のキーワードが含まれている。そこで、キーワード出現密度から検索質問の出現密度を求める。具体的には、キーワード出現密度の各位置での和を求め、検索質問出現密度とする。

⑸適合部分の取り出し

検索質問出現密度を閾値処理することにより、閾値を越える部分を適合部分として出力する。

3.3 処理例

新聞画像を対象に検索実験を行った。実験の詳細は、資料[5]に譲り、ここでは、処理例に基づいて検索の概要を紹介する。

図5に検索対象となる文書画像の1例を示す。この文書画像は1994年度毎日新聞縮刷版を解像度800dpiで取り込んだものであり、紙面右上に減税に関する記事が存在する。

図5　毎日新聞1994年3月6日p.2

図5に対する検索プロセスを図6に示す。この例では、検索質問が単一のキーワード「減税」である。

まず、検索キーへの変換処理の結果、検索キーとしては、「減」と「税」の2つが得られた。ここでは文字に対応する検索キーとなったが、一般には、文字あるいはその一部となる。図6(a), (b)は、それぞれ検索キー「減」が図5に出現する位置、「税」が出現する位置を示す。出現位置とは、検索キーとインデックスのユークリッド距離が一定の閾値以下となった部分画像の中心位置である。

次に、検索キーごとに出現密度(検索キー出現密度)を求めた。結果を図6(c), (d)に示す。ここで、色が濃いほど出現密度が大きい。この図から、検索キー「減」、「税」ともに、画像右上の部分に類似した部分画像が存在することがわかる。

図6(e)は、これらの検索キー出現密度から検索キーワード出現密度を求めた結果である。これを閾値処理することにより、適合部分が図6(f)のように得られる。システムはこの部分を中心として、ユーザに画像を提示する。

図6(g)の正解領域と比較すると、検索によって得られた適合部分は、正解と完全には一致しないが、適切な場所を示す効果はあるといえる。

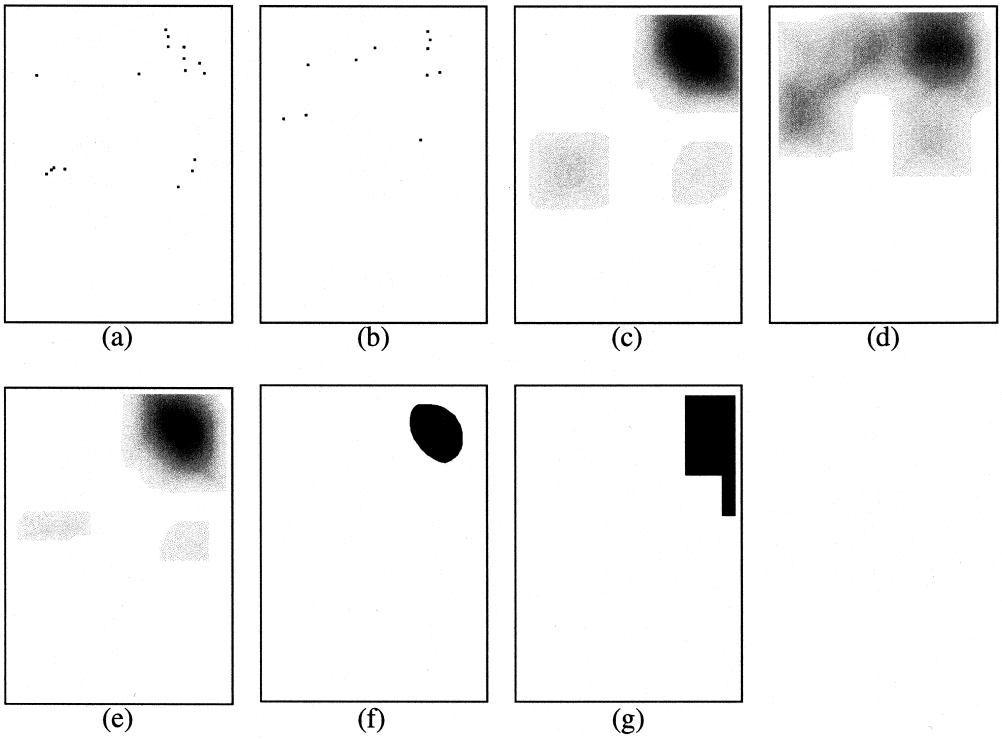

図6　検索プロセス。(a)検索キー「減」の出現、(b)検索キー「税」の出現、(c)(a)から求めた検索キー出現密度、(d)(b)から求めた検索キー出現密度、(e)検索キーワード「減税」の出現密度、(f)閾値処理の結果、(g)正解記事の領域

4　むすび

本報告では、複雑なレイアウトを持つ文書画像に対するレイアウト解析法、文書画像データベースの柔軟な検索法の2つについて述べた。本研究は、文書画像データベースの検索を実現する上での第一ステップであり、実用的な検索を可能とするためには、まだ解決しなければならない問題点が多い。具体的には、検索処理のさらなる高速化に加え、レイアウト解析の研究で得た知見を文書画像データベースの検索法に反映させること、検索法をカラー画像にも対応可能なように拡張することなどが重要な課題である。

謝辞

実験には、「日本語情報検索システム評価用テストコレクションBMIR-J2」、「CD-毎日新聞1994年版」、ならびに「毎日新聞1994年度縮刷版」を利用した。関係各位に感謝する。

参考文献

［1］ K.Kise, M.Iwata and K.Matsumoto, "On the Application of Voronoi Diagrams to Page Segmentation", Proc. of Workshop on Document Layout Interpretation and its Applications (DLIA99) , IV-C, pp.1–4 (1999).

［2］ K.Kise, A.Sato and M.Iwata, "Segmentation of Page Images Using the Area Voronoi Diagram", Computer Vision and Image Understanding, Vol.70, No.3, pp.370–382 (1998).

［3］ K.Kise, M.Iwata, A.Dengel and K.Matsumoto, "A Computational Geometric Approach to Text-line Extraction from Binary Document Images", Proc. of the Third IAPR Workshop on Document Analysis Systems (DAS98), pp.346–355 (1998).

［4］ 岩田基、黄瀬浩一、松本啓之亮、"隣接グラフを用いた欧文文書画像からの文字列抽出"、情報処理学会論文誌、Vol.40, No.8, pp.3239–3248(1999).

［5］辻野雅章、黄瀬浩一、松本啓之亮、"特徴ベクトルの出現密度分布に基づく印刷文書画像の部分検索"、電子情報通信学会 パターン認識・メディア理解研究会、PRMU99-226 (2000).
［6］辻野雅章、黄瀬浩一、松本啓之亮、"キーワード出現密度分布に基づく文書画像の部分検索"、電子情報通信学会 2000年総合大会、D-12-33 (2000).
［7］K.Kise, H.Mizuno, M.Yamaguchi and K.Matsumoto, "On the Use of Density Distributions of Keywords for Automated Generation of Hypertext Links from Arbitrary Parts of Documents", Proc. of the 5th International Conference on Document Analysis and Recognition (ICDAR' 99), pp.301–304 (1999).
［8］水野浩之、黄瀬浩一、松本啓之亮、"単語の出現密度と偏出度を用いた図表と説明テキストの対応付け"、情報処理学会論文誌、Vol.40, No.12, pp.4400–4403 (1999).

＜発 表 資 料＞

題　　名	掲載誌・学会名等	発表年月
A Computional Geometric Approach to Textline Extraction from Binary Document Images	Proc. of the Third IAPR Workshop on Documennt Analysis Systems (DAS'98)	1998年11月
隣接グラフを用いた欧文文書画像からの文字列抽出	情報処理学会論文誌	1999年8月
On the Application of Voronoi Diagrams to Page Segmenntation	Proc. of Worhshop on Document Layout Interpretation and its Applications (DLIA99)	1999年9月
On the Use of Density Distributions of Keywords for Automated Generation of Hypertext Links	Proc. of the 5th Int'l Conf. on Document	1999年9月
from Arbitrary Parts of Documents	Analysis and Recogniton (ICDAR'99)	0000年0月
特徴ベクトルの出現密度分布に基づく印刷文書画像の部分検索	電子情報通信学会 パターン認識・メディア理解研究会	2000年2月
キーワード出現密度分布に基づく文書画像の部分検索	電子情報通信学会 2000年総合大会	2000年3月