コンピュータビジョン　52—13
（1988. 1. 22）

文書画像理解による記事の自動抽出

Article Extraction by Document Understanding

辻本　修一　　麻田　治男
Shuich Tsujimoto　　Haruo Asada

㈱東芝　総合研究所
Toshiba corporation, Research and Development Center

あらまし

文書構造は幾何学的構造と論理的構造の２つの構造で表現される。幾何学的構造とは隣接文字列より構成される部分領域の配置関係（レイアウト）を記述するものであり、論理的構造とは部分領域間の意味的な係わり関係を記述するものである。本来、この係わり関係を完全に得るためにはその内容を解析する必要がある。しかし、多くの文書では、その係わり関係を一目で了解出来るようにブロックがレイアウトされている。よって、本論文では文書の幾何学的構造よりその論理的構造を獲得することを考え、それを簡単なアルゴリズムで実現する。

ABSTRACT Document structure is represented in two ways: a geometric structure to represent the configuration between blocks which contain neighboring text lines, and a logical structure to represent the semantics of the blocks. We consider "document understanding" as obtaining a correspondence between the two. In this paper, we present an algorithm to describe the correspondence, and propose a method for extracting each article from a document and recognizing the text parts in the article. Moreover, the method is applied to generally published documents with various kinds of formats and proves to be available and effective in actual fields.

1. はじめに

近年、新聞、雑誌、公文書等の印刷文書の構造を理解し、その内容を読む技術の構築が望まれている。また、文書の構造を理解する際にその部分領域間の意味的な支配関係をも記述出来れば、記事の構成要素単位の処理が可能となるため、ドキュメント管理やデータベース入力、電子出版において更に多様な処理を実現出来る。本論文では、多記事から構成される、書式が未知の英文文書の構造を理解し、各記事を抽出する手法について考察する。

上記要求を実現するためには、入力画像を互いに性質の異なる部分領域に分割する構造解析[5]と、各部分領域間の関係（つながり、包含、並列等）を調べる構造理解の２つのプロセスが必要となる。

従来、構造解析に関してはランレングス[1]、周辺分布[4,6]、縮小画像[8]、拡大、縮退[9]を用いる手法が報告されている。しかし、これらの手法には、入り組んだ記事を含む文書を取り扱う事が出来なかったり[4,6]、多くの処理時間を要したり[1,8,9]するという問題があった。また、構造理解に関しては対象を、明確な規則に従ってレイアウトされている文書に限定した研究[2,7]が多く、書式の多様性という点において満足出来るものではなかった。この他に書式を予め言語で定義しておき、それに基づいて文書構造を理解していく手法[10]も提案されているが、各文書ごとに書式定義を行なう必要があった。対象を限定しないものとしては、文書の配置関係を文書画像の配置構造より生成する手法[3]が提案されているが、周辺分布を用いて構造解析を行なっているため、入り組んだ記事を取り扱うことが出来なかった。

ここで、文書構造について考えてみる。文書構造は幾何学的構造と論理的構造の２つの構造で表現される。幾何学的構造とは隣接文字列より構成される部分領域の配置関係を記述するものであり、これはレイアウトに相当する。また、論理的構造とは部分領域間の意味的な係わりを記述するものであり、例えば、１つの記事に対するタイトルブロックとか、そ

のタイトルブロックに対する本文ブロックとかいう関係を記述する。本論文においては、入力画像より幾何学的構造を得ることを構造解析、幾何学的構造より論理的構造を生成することを構造理解と定義する。ブロック間の意味的な係わりというのは本来、その内容のつながり関係を調べることにより得られる。しかし、多くの文書では、視覚効果を考慮して、その係わり関係を一目で了解出来るようにブロックがレイアウトされている。逆の見方をすれば、レイアウト、即ち、幾何学的構造よりその論理的構造を作り上げることが出来るはずである。これが我々の着目点であり、本論文においては、簡単なアルゴリズムで幾何学的構造より論理的構造を求めることが出来ることを示す。また、構造解析においても、ランレングス符号による画像処理を積極的に利用した新しい手法を提案する。ランレングス表現はデータ圧縮のみならず計算量の圧縮にも効果がある。

本論文では、まず、2つの文書構造を定義し、幾何学的構造より論理的構造を記述するアルゴリズムを提案する。次に、入力画像より幾何学的構造を作成する手法について述べる。最後に、本手法の有効性を実験により確認する。

2. 文書の階層的構造

2. 1 幾何学的構造

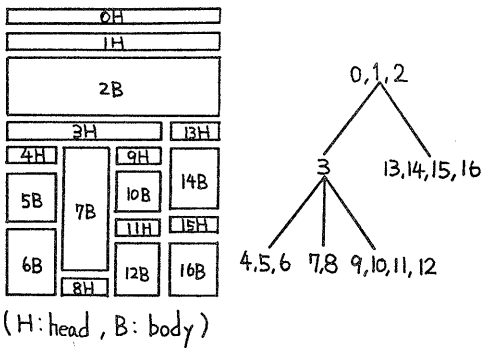
文書は幾何学的な構造をもっている。文書は文字列より構成されていて、隣接する文字列はひとつのブロックにまとめられる。ブロックはパラグラフの様な複数のサブブロックに分割される場合もある。この際、ブロックの配置関係は木を用いて表現される。つまり、木の各ノードはブロックに対応し、それはサブブロックのリストで表現される。また、段組みはノードの広がりに対応する。この様な文書構造を幾何学的構造と呼ぶ。図1にその一例を示す。 (a)で表現されるサブブロックの幾何学的構造の木表現が (b)である。

ここで、サブブロックをその物理的性質により2つの項目に分類する。1つは"head"項目で、これは、数行の文字列を含み、数段にまたがっていたり、センタリング処理が施されているサブブロックに対して与えられる。もう1つは"body"項目で、これは、"head"項目でないサブブロックに対して与えられる（図1参照）。

2. 2 論理的構造

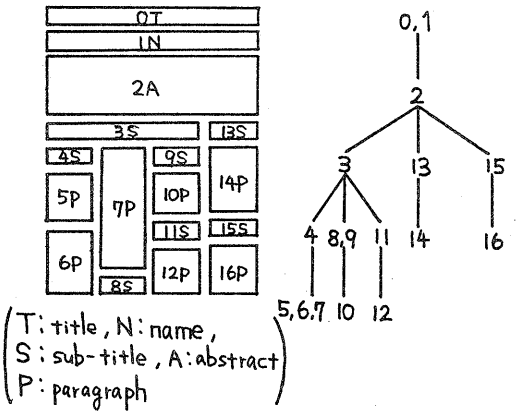
前節では、各サブブロックをその物理的性質という立場において、2つの項目に分類した。その文書の内容（中身）に関して言えば、サブブロックはタイトル、サブタイトル、アブストラクト、パラグラフ等の意味に分類され、これはサブブロックの内容と他のサブブロックとの配置関係が調べられて始めて決定される。

各サブブロックの意味とその意味的な係わり、つまり、意味的な支配関係は木を用いて表現され、それを論理的構造と呼ぶ。同じ意味を持つサブブロックの集合は各ノードに対応



(a) document divided into geometric sub-block (b) geometric structure represented by a tree

Fig.1 Geometric structure.



(a) document divided into logical sub-block (b) logical structure represented by a tree

Fig.2 Logical structure.

し、意味的な支配関係は木の深さに対応する。図2にその一例を示す。図中(b) は (a)で表現されるサブブロックの論理的構造を木表現したものである。

2. 3 構造理解

一般に、異なる意味を持つサブブロックはそのことを読者が一目で了解出来る様にレイアウトされているのが普通である。つまり、サブブロックの意味の違いをそのレイアウトに反映させることによって、読者にそのことを強調している。これは結局、幾何学的構造を解析するだけで、論理的構造を獲得出来ると言うことになる。

論理的構造から幾何学的構造への対応は1対多対応であり、

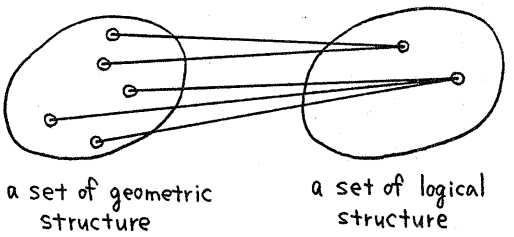


Fig.3 Multiple geometric structure corresponding to a single logical structure.

これはレイアウト変換の多様性に相当する。これに対し、その逆である幾何学的構造から論理的構造への対応は一意に決定出来るはずであるというのが本論文の主張であるが（図3参照）、本節においては、これを決定するアルゴリズムを提案する。論理的構造は下に示す4つの操作（ステップa－d）を幾何学的構造に施すことにより得られる。この操作の一例を図4に示す。なお、これらの処理に先だって、幾何学的構造の木の各ノードには縦型の順序に従って番号が付けられており（子は弟よりも先である）、これは、読みの順序に対応している。

ステップa

リストの先頭が"body"であるターミナルノードAに対しては、その1つ手前のノードBがターミナルノードであればノードBのリストの終端に"body"サブブロックを移動させる。

これは、ブロックの先頭サブブロックが"body"である場合に対する操作である。

ステップb

リストの終端が"head"であるターミナルノードBに対しては、その次のノードAがターミナルノードであればノードAのリストの先頭サブブロックを今、着目している"head"サブブロックに付け加える。

これは、ブロックの終端サブブロックが"head"である場合に対する操作である。

ステップc

1つ以上のサブブロックと1つの "head-body"列（1つ以上の"head"と1つ以上の"body"から成る）で構成されるノードAに対しては、1つの "head-body"列で構成されるノードDを新たに作り、それを弟として付け加える。但し、この際、ノードAが子孫ノード群Cを持っていたらそれも一緒に移動させる。また、ノードAに親がいなければNULLの親ノードを作り、上述した処理を行う。

これは、1つのノードが1つの意味を表す様にするための前処理操作である。

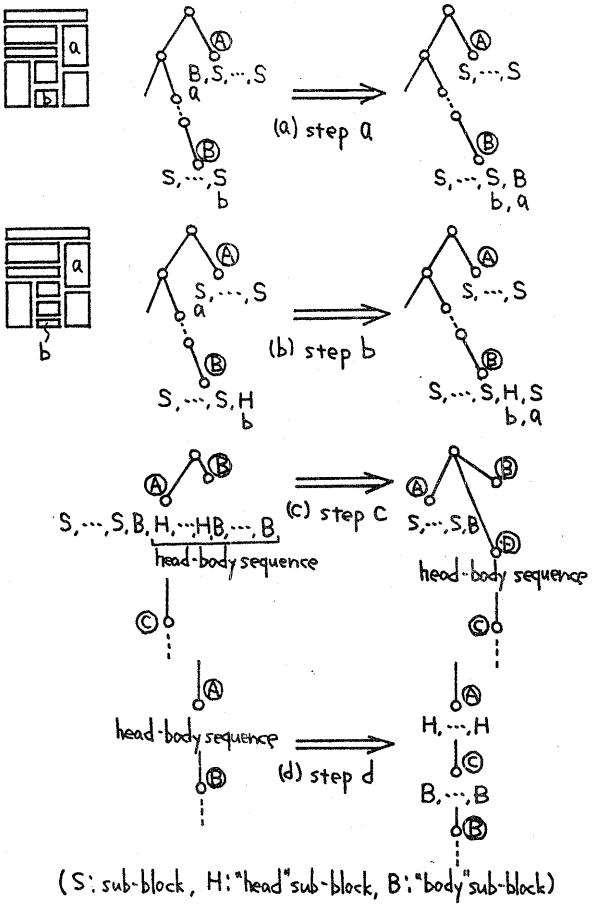


Fig.4 Converting a geometric structure to a logical structure.

ステップd

1つの "head-body"列で構成されるノードAに対しては、"body"列で構成されるノードCを新たに作り、それを子供として付け加える。

これは、1つのノードが1つの意味を表す様にするための操作である。

上述した4つの操作により、幾何学的構造（図1(b)）から論理的構造（図2(b)）が作られる一例を図5に示す。

ここで、サブブロックの項目名はより情報量の多いものにすることが出来る。ルートノードにある"head"サブブロックは" タイトル" や" 著者名" 等を表し、その他のノードにある"head"サブブロックは" サブタイトル" を表すものと決定される。また、ターミナルノードにある"body"サブブロックは" パラグラフ" を表し、その他のノードにある"body"サブブロックは" アブストラクト" 等を表すものと決定される。

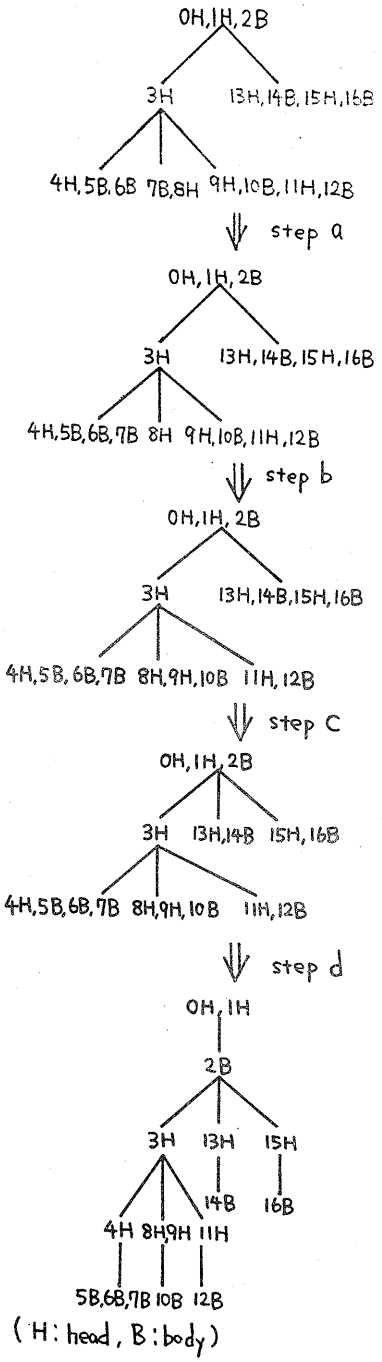


Fig.5 An example converting geometric tree to logical tree.

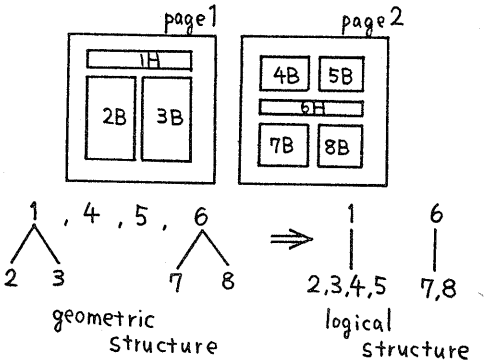


Fig.6 Multi-article corresponding to multi-tree.

2．4　多記事の取り扱い

多記事を含む文書の意味的構造は木のリストで表現される。ルートノードが"head"サブブロックである木は１つの記事を表し、ルートノードが"head"サブブロックでない木はそれの１つ手前の木に付け加えられる（図６参照）。これにより、複雑に入り組んだ文書でもその意味的構造を表現する木のルートノードを調べるだけで各記事を抽出することが出来る。

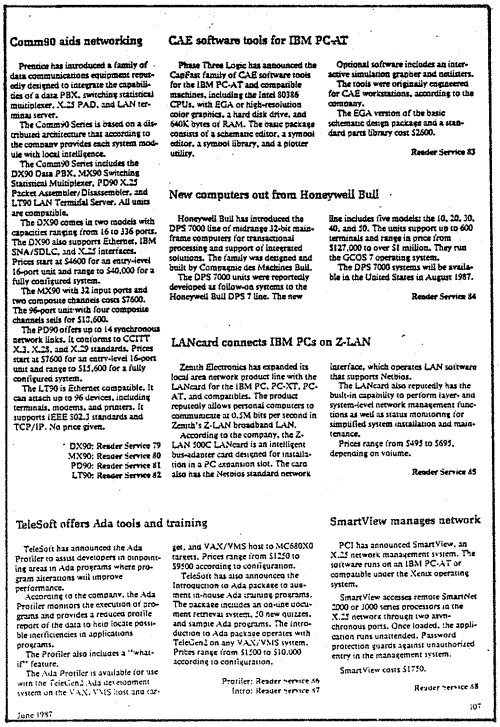
3．構造解析

3．1　フィールド抽出

本論文では、文字列、あるいはその一部を取り囲む最小矩形をフィールドと定義し、本節ではこの抽出法について述べる。また、画像の表現法としてはランレングス表現を用いるが、これはデータ圧縮のみならず、処理量の圧縮にも効果がある。

文書画像処理におけるフィールド抽出は従来、次に示す手法が用いられてきた。まず、入力画像を水平、垂直方向のランで表現し、短い白ランを黒ランに置き換える操作をそれぞれのランに適用する。そして、それらをいったん、ビットマップに展開し、そのAND をとった画像を水平方向のランで表現する。最後に再び短い白ランを黒ランに置き換える、というものであった。この手法は有効ではあるが、ビットマップデータとランデータを共に用いる必要があったり、処理量が多い等、実用的ではなかった。

これに対し、本論文では、水平ランのみについて、短い白ランを黒ランに置き換え、ラベリング演算を施し、そして、同じラベルをもつランを取り囲む最小矩形をフィールドとして抽出する。この段階では、１つのフィールドが１つの文字列を代表するとは限らないが、文字列は次節で示すように段組みを決定してから抽出することにする。この処理により、効率の良い文字列抽出が可能となる。図７に示す入力文書画像（"Computer", June, 1987より引用）のフィールド抽出結果を図８に示す。図中の矢印、数字については後述する。な



Comm90 aids networking

Prentice has introduced a family of data communications equipment reportedly designed to integrate the capabilities of a data PBX, switching statistical multiplexer, X.25 PAD, and LAN terminal server.

The Comm90 Series is based on a distributed architecture that according to the company provides each system module with local intelligence.

The Comm90 Series includes the DX90 Data PBX, MX90 Switching Statistical Multiplexer, PD90 X.25 Packet Assembler/Disassembler, and LT90 LAN Terminal Server. All units are compatible.

The DX90 comes in two models with capacities ranging from 16 to 336 ports. The DX90 also supports Ethernet, IBM SNA/SDLC, and X.25 interfaces. Prices start at $4600 for an entry-level 16-port unit and range to $40,000 for a fully configured system.

The MX90 with 24 input ports and two composite channels costs $7600. The 96-port unit with four composite channels sells for $12,600.

The PD90 offers up to 14 synchronous network links. It conforms to CCITT X.3, X.28, and X.29 standards. Prices start at $7600 for an entry-level 16-port unit and range to $15,600 for a fully configured system.

The LT90 is Ethernet compatible. It can attach up to 96 devices, including terminals, modems, and printers. It supports IEEE 802.3 standards and TCP/IP. No price given.

DX90: Reader Service 79
MX90: Reader Service 80
PD90: Reader Service 81
LT90: Reader Service 82

CAE software tools for IBM PC-AT

Please Three Logic has announced the CanFast family of CAE software tools for the IBM PC-AT and compatible machines, including the Intel 80386 CPUs, with EGA or high-resolution color graphics, a hard disk drive, and 640K bytes of RAM. The basic package consists of a schematic editor, a symbol editor, a symbol library, and a plotter utility.

Optional software includes an interactive simulation grapher and netlisters.

The tools were originally engineered for CAE workstations, according to the company.

The EGA version of the basic schematic design package and a standard parts library cost $2600.

Reader Service 83

New computers out from Honeywell Bull

Honeywell Bull has introduced the DPS 7000 line of midrange 32-bit mainframe computers for transactional processing and support of integrated solutions. The family was designed and built by Compagnie des Machines Bull.

The DPS 7000 units were reportedly developed as follow-on systems to the Honeywell Bull DPS 7 line. The new line includes five models: the 10, 20, 30, 40, and 50. The units support up to 600 terminals and range in price from $127,000 to over $1 million. They run the GCOS 7 operating system.

The DPS 7000 systems will be available in the United States in August 1987.

Reader Service 84

LANcard connects IBM PCs on Z-LAN

Zenith Electronics has expanded its local area network product line with the LANcard for the IBM PC, PC-XT, PC-AT, and compatibles. The product reputedly allows personal computers to communicate at 0.5M bits per second in Zenith's Z-LAN broadband LAN.

According to the company, the Z-LAN 500C LANcard is an intelligent bus-adapter card designed for installation in a PC expansion slot. The card also has the Netbios standard network interface, which operates LAN software that supports Netbios.

The LANcard also reputedly has the built-in capability to perform layer- and system-level network management functions as well as status monitoring for simplified system installation and maintenance.

Prices range from $495 to $695, depending on volume.

Reader Service 85

TeleSoft offers Ada tools and training

TeleSoft has announced the Ada Profiler to assist developers in pinpointing areas in Ada programs where program alterations will improve performance.

According to the company, the Ada Profiler monitors the execution of programs and provides a reduced profile report of the data to help locate possible inefficiencies in applications programs.

The Profiler also includes a "what-if" feature.

The Ada Profiler is available for use with the TeleUse2 Ada development system on the VAX/VMS host and target, and VAX/VMS host to MC680X0 targets. Prices range from $1250 to $9500 according to configuration.

TeleSoft has also announced the Introduction to Ada package to augment in-house Ada training programs. The package includes an on-site document retrieval system, 20 new quizzes, and sample Ada programs. The introduction to Ada package operates with TeleUse2 on any VAX/VMS system. Prices range from $1500 to $10,000 according to configuration.

Profiler: Reader service 86
Introl: Reader service 87

SmartView manages network

PCI has announced SmartView, an X.25 network management system. The software runs on an IBM PC-AT or compatible under the Xenix operating system.

SmartView accesses remote SmartNet 2000 or 3000 series processors in the X.25 network through two asynchronous ports. Once loaded, the application runs unattended. Password protection guards against unauthorized entry in the management system.

SmartView costs $1750.

Reader service 88

June 1987

107

Fig.7 Document image.

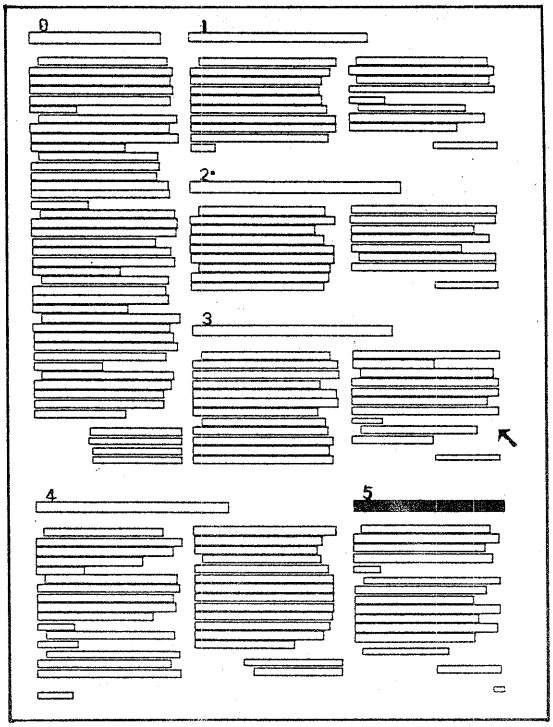


Fig.8 Field extraction.

お、ここでのラベリング演算を含む基本画像処理は全てランを処理の単位として行なわれている。この段階でフィールドはいくつかのカテゴリに分類される。例えば、細長いフィールドはアンダーラインやフィールドセパレータ等を代表する直線成分とみなされる。

3．2　ブロック抽出

本節では、同一段組みにある隣接フィールドをブロックに統合する。

段組み位置は各フィールドの左端点を垂直方向に射影し、そのピークを調べることにより、決定される（図9(a) 参照）。

各フィールドと段組みとの係わり関係はグループ番号により記述する（図9(b) 参照）。グループ番号は次の式により求められる。

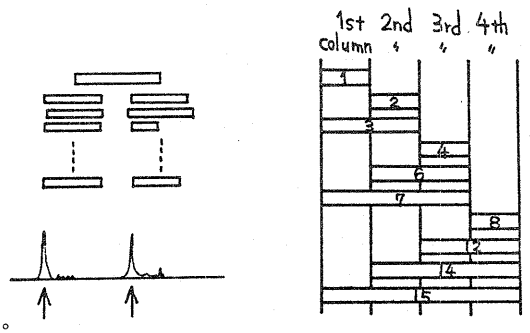
$$\text{グループ番号} = \sum_{i=1}^{N} cp[i] * 2^{(i-1)}$$

cp[i] = 1 （フィールドがｉ番目の段に含まれている時）

cp[i] = 0 （それ以外）

Ｎは段組みの数である。

同じグループ番号をもつ隣接フィールドがブロックに統合される。この時、フィールドセパレータや大きな空白があればそこでもブロックに区切られる。また、この段階でフィールドを文字列に統合出来る。



(a) how to determine the column position　(b) definition of group number

Fig.9 Column setting.

3．3　サブブロック抽出

ブロックはフィールド間距離が基本フィールド間距離よりも大きい位置やインデントが存在する位置でサブブロックに分割される。基本フィールド間距離はブロック抽出の際、その中のフィールドの値より決定される。

各サブブロックには２．１節に示した方法によって、項目名"head"と"body"が与えられる。

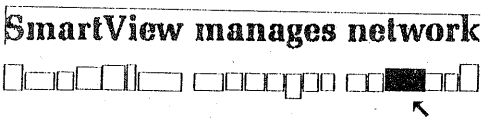
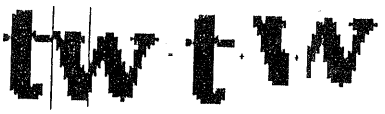


Fig.10 Rectangles representing characters.

(a) in case of 2 characters



(b) in case of 3 characters

Fig.11 Segmentation for the connected characters.

3．4　幾何学的構造の生成

ブロックは幾何学的構造を記述する木の各ノードに対応しており、ブロックAがブロックBの上方に位置していて、ブロックAがブロックBのグループ番号より大きなグループ番号をもっているのならブロックAに対応するノードはブロックBに対応するノードの親となる。

ノードはサブブロックのリストで表現されるが、物理的に上方に位置づけられているサブブロックがリストの前方になるように順序づけられる。

文書がいくつかの記事から構成されている場合、少なくともその数の分だけの木が生成されるが、そのリストの順番は読み順が左上には戻らないという原則に従って決定される。

4．実験

4．1　論理的構造の生成

第2章で提案したアルゴリズムより、論理的構造が幾何学的構造から生成される。そして、各木では縦型の順序により読み順が決定される。図8中の数字は記事の読み順を示している。

4．2　文字検切

各フィールド内の画像データにラベリング処理を施すことにより、連結領域を抽出する。そして、それを取り囲む最小矩形を得る。多くの場合、この連結領域は1文字を表わしているが、"i"や"j"のドットように二つの連結領域に分離される場合や何文字かが接触してひとつの連結領域になることもある。

ある小さな最小矩形が他の最小矩形の上方に位置していたら、それをマージして1文字として扱う。図10（図8の黒く塗りつぶされたフィールド）の"View"の"i"等がその一例である。

図10の"SmartView"の"rt"や"ew"、及び、"netwotk"の"tw"等の接触文字に対しては、接触文字数を推定し、その文字数に対する検切を行なう。接触文字数として基本文字幅と着目矩形の幅との比率より幾つかの推定値が挙げられる。なお、基本文字幅はサブブロック単位に、すべての最小矩形が抽出され、その幅の中央値を選んでいる。

各推定文字数に対する最も適切な検切位置は次の手順で決定する。（1）　推定文字数で矩形の幅を均等に分割した位置の周辺にサーチ区間を設定する。（2）　連結領域の輪郭が凹の位置を検切候補位置として設定する。（3）　各文字のサーチ区間における検切候補位置のうち、その切断量（切断線と文字パターンとの交線の長さの総和）が最も小さい位置を検切位置と決定する。図11は図10の"network"の"tw"の検切の例で、(a)は接触文字数を2文字と推定した場合、(b)は3文字と推定した場合の例である。

最終的な検切位置は次節で述べる文字認識、単語照合処理の後、決定される。なお、輪郭が凹となる位置を検出するのにはランレングス表現が非常に適している。

4．3　文字認識と単語照合

切り出された各文字に対して複合類似度[11]による文字認識が行われ、幾通りかの候補が挙げられる。各文字は文字間スペースやコンマ、ピリオド、ハイフン等を参考にして、単語に統合される。そして、検切の各候補、文字認識の各候補に対して、単語照合が行なわれ、最も適切なものが選ばれる。また、文字認識はマルチフォント対応である。

4．4　実験結果

本論文で提案した手法を雑誌、論文、新聞、書籍の数種の英文文書に適用した結果、各記事をその読み順通りに正確に抽出出来ることが分かった。例えば、図8中のカーソルで指し示した記事の読み取り結果を図12に示すが、これからも記事が正確に抽出され、それが正しい順序で読まれていることが確認出来る。この文書には約4200文字含まれているが、この場合の文字認識率は99%であった。なお、この実験で用いた文書の文字フォントは文字認識辞書を作成した時に用いたフォントとは異なっていたということに注意しておく。また、処理をすべてソフトウェアで行なった場合、処理時間はSUN3ワークステーションで約4分半であった。その内訳は、スキャンに20秒、ビットマップからランデータへの変換に 7秒、構造解析に10秒、構造理解に 1秒、検切、文字認識、単語照合に約 4分であり、この処理速度は15（文字/sec）に相当するものである。

[article_no.3]
LANcard connects IBM PCs on Z-LAn
Zenith Electronics has expanded its
local area network product line with the
LANcard for the IBM PC, PC-XT,
PC-AT, and compatibles. The product
reputedly allows personal computers to
communicate at 0.5M bits per second in
Zenith's Z-LAN broadband LAN.
According to the company, the
ZLAN 500C LANcard is an intelligent
bus-adapter card designed for
installation in a PC expansion slot. The card
also has the Netbios standard network
interface, which operates LAN software
that supports Netbios.
The LANcard also reputedly has the
built-in capability to perform layer- and
system-level network management
functions as well as status monitoring for
simplified system installation and
maintenance.
Prices range from $495 to $695,
depending on volume.
Reader Service 85

Fig.12 Recognizing characters.

5. むすび

本論文では、文書構造を、レイアウトを記述する幾何学的構造と文書のつながりを記述する論理的構造の２つの構造で表現した。そして、幾何学的構造から論理的構造を作成することを考え、そのアルゴリズムを提案した。また、ランレングスのまま、画像処理を行なうことによりすべての処理をソフトウェアで実現することを可能にし、システムをコンパクトで、ハードウェアに依存しないものにすることが出来た。実験により、多段組、多記事、多種多様なフォントを含む文書に対しても、各記事を正確に抽出し、実用的な処理時間で、かつ、高い文字認識率で読みとることが出来ることを確認した。

現在、本システムの構造解析部は縦のフィールドセパレータや表を扱うことが出来ない。また、構造理解部は図表のキャプションやページ番号等の文書以外部分と文書部分との係わりを記述することが出来ない。今後はこれらの欠点を取り除いて、入力文書に対する制限を出来るだけ少なくしていく予定である。

参考文献

[1] F.M.Wahl, K.Y.Wong and R.G.Casey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents", Computer Graphics and Image Processing 20, pp. 375-390,1982.

[2] 山田、西村、野口、豊田、" 記事の形状に着目した英字新聞の領域分割" 、情処全大2B-4,1983.

[3] 辻、 " 文書画像からの構造記述の生成" 、情処全大5K-8,1987.

[4] 辻、浅井、" スプリット検出法に基づく頁画像の構造解析" 、信学技報,PRL85-17(1985).

[5] S. N. Srihari and G. W. Zack, "Document Image Analysis", Proc. 8th ICPR, Paris, pp. 434-436, 1986.

[6] 秋山、増田、" 周辺分布、線密度、外接矩形特徴を併用した文書画像の領域分割" 、 信学論(D), J69-D 、 8、pp.1987-1996(1986).

[7] K.Inagaki and T.Kato, "MACSYM: A Hierarchical Parallel Image Processing System for Event-Driven Pattern Understanding of Documents", Pattern Recognition, Vol. 17, No. 1,pp. 85-108 1984.

[8] Y.Maeda, F.Yoda, K.Matsuura and H.Namba, "Character Segmentationin Japanese Hand-written Document Images", Proc.8th ICPR, Paris, pp.769-772, 1986.

[9] 中村、鈴木、南、" 横書き日本語文書における個別文字の抽出" 、信学論(D),J68-D, 11, pp.1899-1909, 1985.

[10] J.Higashino, H.Fujisawa, Y.Nakano and M. Ejiri, "A Knowledge-based Segmentation Method for Document Understanding", Proc. 8th ICPR, Paris, pp. 745-748, 1986.

[11] T.Iijima, H.Genchi and K.Mori, "A Theory of Character Recognition by Pattern Matching Method", Proc. 1st International Conference on Pattern Recognition, pp. 50-56,1973.