情報処理学会第37回(昭和63年後期)全国大会

複雑な構造を持つ表の認識に関する基礎検討

6W-8

児島 治彦　清末 悌之　秋山 照雄

NTTヒューマンインタフェース研究所

1．まえがき

文書画像から切り出された図表領域を対象として、図表の物理構造・論理構造の認識法について検討している。本稿では複雑な構造を持つ表の認識について基礎検討を行った結果について報告する。はじめに既存文書中の表の調査で得られた構造の特徴と、構造からみた表の分類について述べる。次に表構造認識の全体の流れを述べた上で、調査で得られた『罫線の両端は直角方向の罫線と接する』という知見をもとに、表から罫線と罫線に囲まれたブロックを再帰的に抽出する手法を提案する。基礎実験により、本手法の有効性を確認した。

2．表の調査と分類

表構造認識の指針を得るため、既存文書中の表について調査を行い、物理構造の特徴を抽出し、構造からみた表の分類を行った。対象は和文文書として情処論文誌4冊、信学論3冊、研究実用化報告5冊、NTT施設3冊、英文文書としてProceedings of IEEE5冊、IEEE Trans. ASSP5冊、Review of the ECL 3冊、さらにワープロ作成文書をオフセット印刷した情処全大予稿集1冊、以上の文書中の表で、全部で650編である。物理構造として特に罫線に着目して特徴を抽出し、罫線の有無、階層構造の有無などを用いて表の分類を行った。得られた知見の主なものを以下にあげる。ここで外枠の縦罫線が存在する表をCLOSE型、存在しない表をOPEN型と呼ぶ。

a)表中の図形は主に水平・垂直線（罫線）で構成される
b)文字と図形は接触しない
c)かすれ、破線などにより罫線の連続性は保障されない
d)罫線幅に比べ、平均的な文字サイズは十分大きい
e)罫線の両端は直角方向の罫線と接する（OPEN型の表は外枠縦罫線の存在を仮定）
f)組版作成の表では組版の複雑さの関係からOPEN型の表が多く（予稿集以外の文書ではOPEN：CLOSE = 8：2）、特に外国文書では縦罫線がほとんど存在しない（但し、オフセット印刷文書中の表には縦罫線が存在する）
g)1つの枠内の文字列がすべて1行で済む場合は、横罫線が省略されることが多い
h)斜め線をもつ表は全体の5％で、最左上の枠内に存在する（英文文書ではほとんど存在しない）
i)カテゴリ（項目）内で階層性を有する表（階層構造を持つ表と呼ぶ）は全体の28．6％
j)罫線が一部除去された表は全体の2％
k)図（罫線・斜め線以外の図形）を含む表は3．4％

ここでは分類されたもののうち、図を含まない表を対象として認識手法を検討する。

3．全体の処理の流れ

表認識処理の全体の流れを図1に示す。文書画像から抽出された表領域について、傾き補正の後、罫線と罫線に囲まれたブロックの抽出などを行い、抽出された各ブロックの内部から文字列を抽出する。同一ブロックに複数行の横文字列があるとき、各行の最右文字位置などのレイアウト情報をもとに横罫線が省略されているか否かを判定し、省略されていれば該当位置に横罫線を追加する。抽出された文字列について、文字切出しと文字認識を行う。認識結果について言語処理を行い、横罫線の追加判定、ブロック間の意味的な関係情報抽出を行う。最後に、縦方向・横方向のブロック間関係記述、階層構造の記述を行う。構造まで記述することにより、表の再加工処理が容易になる[1]。ここでは、このうちの罫線およびブロック抽出アルゴリズムについて報告する。

4．罫線およびブロック抽出アルゴリズム

『罫線の両端は直角方向の罫線と接する』ことに着目した、再帰的手法による罫線およびブロック抽出アルゴリズムについて述べる。表調査で得られた知見のうち、a)～e)を前提として本アルゴリズムを作成した。なお、事前に表画像の傾きが正規化されていることが本アルゴリズム適用の条件となる。

知見a)～d)により、周辺分布は罫線抽出に有効である。

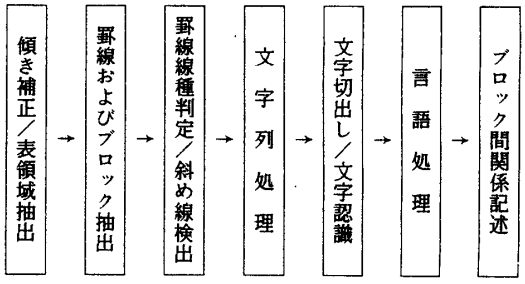

図1　表認識処理の全体の流れ

A Study on Table Recognition with Complex Structure
Haruhiko KOJIMA, Yasuyuki KIYOSUE and Teruo AKIYAMA
NTT Human Interface Laboratories

周辺分布のヒストグラムから、ある閾値以上の高さ、ある閾値以下の幅を持つ山を求めることにより、罫線位置を推定できる[2]。ランを用いた手法では、破線、かすれへの対処が問題となる[3]。ただし、表領域全体に対して周辺分布をとるだけでは複雑な構造の表からあまねく線分を抽出することはできない。また、線分の端点位置を求めるのに、単純なマスク処理では多くの時間を要する。以上の問題を解決するため、知見e)を用いたアルゴリズムを提案する。以下にアルゴリズムの概要を示す。

1)表領域全体で周辺分布をとり、罫線候補を抽出する。
2)外枠罫線を決定する。OPEN型の場合、便宜的に外枠縦罫線を追加する。
3)罫線候補のうち、両端が外枠罫線と接するものを罫線として抽出する。
4)抽出された罫線で構成されるブロックを求め、各ブロック領域内で周辺分布をとり、罫線候補を抽出する。
5)抽出された罫線候補のうち、両端が該ブロックの辺と接するものを罫線として抽出する。
6)抽出された罫線で構成されるブロックを求め、あらたに得られた各ブロック領域内で周辺分布をとり、罫線候補を抽出する。罫線候補があらたに抽出された場合、5)に戻る。候補がなくなるまで再帰的処理を行う。

□

両端が他の罫線と接触しない罫線候補は、候補から除かれる。上述の知見a)～d)は手法1),4),6)に、e)は3),5)に反映されている。本手法の適用例を図2に示す。

5．基礎実験

本手法の有効性を確認するため、少数データによる基礎実験を行った。対象は以下に示す6編で、分類された各クラスから代表例を用いた。

1－CLOSE型、罫線が一部除去された和文の表(11X8.5cm)
2－OPEN型、階層構造を持つ和文の表(11X15cm)
3－OPEN型、階層構造を持つ和文の表(7X8.5cm)
4－OPEN型、罫線が一部除去された英文の表(14X8.5cm)
5－OPEN型、階層構造を持つ英文の表(16.5X13cm)
6－CLOSE型、階層構造を持つ和文の表(7X10cm)

出典は1,6がＮＴＴ施設、2,3が研究実用化報告、4,5がReview of the ECLである。各データは１６本／mmのＦＡＸを用いて２値入力された画像で、予めＬＰＰ法で傾きを正規化してある[4]。罫線抽出時は、実験装置のメモリサイズの関係から画像を３．２倍（５本／mm）に縮小した。使用データのうち、4を図３に示す。

本実験では、６編いずれも罫線およびブロックを正しく抽出できた。印刷文書中の表では文字と図形が接触しない、などの条件が本手法の有効性に寄与していると考える。しかし、本手法を多様なデータに適用するためには、細くかすれた線や太線に対処可能な罫線候補抽出パラメータの最適自動設定法について検討する必要がある。

6．おわりに

複雑な構造を持つ表の認識について基礎検討を行った。既存文書中の表の調査を行い、表構造の特徴抽出・分類を行った。表構造認識の全体の流れを述べた上で、『罫線の両端は直角方向の罫線と接する』という知見を用いて罫線とブロックを再帰的に抽出する手法を提案し、基礎実験により有効性を確認した。今後の課題には、罫線候補抽出パラメータ最適自動設定法、大量データによる検証、さらにブロック中の文字列処理やブロック間構造記述の検討がある。

参考文献

[1] 清末他：再加工を考慮した表のデータ構造に関する一考察，情処37回全大投稿予定（1988).
[2] 佐藤他：文書入力のための表構造の認識，昭63信学春季全大，2-224（1988).
[3] 宮原：文書読取りにおける線分処理の検討，昭62信学部門別全大，1-78（1987).
[4] 秋山他：書式指定情報によらない紙面構成要素抽出法，信学論，Vol.J66D，No.1（1983).

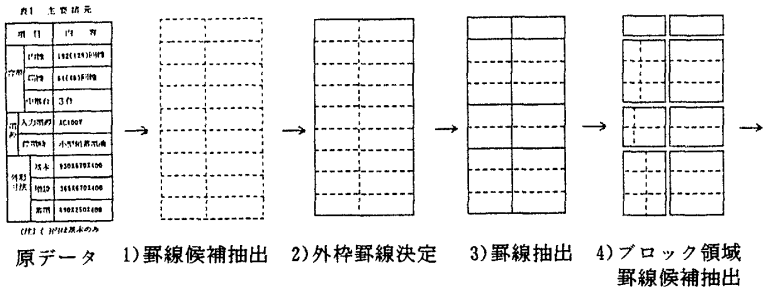

図２　罫線およびブロック抽出アルゴリズム適用例

Table 1 IKTS SYSTEM DIMENSIONS.

Items	Small Size Model	Medium Size Model	Large Size Model
Port Capacity (co line/station)	4/20	12/40	48/160
Outside Line Type	Analog subscriber line, tie lines (loop/disconnect, ring-down) dial-in line, facsimile-net line		
Service Trunks	Pooled modem, pooled speakerphone circuit, voice storage/synthesis, co-co transfer circuit, conference circuit, SMDR		
Terminal Types	Key telephone set, direct station selection console, office line display, single-line telephone, personal computer, paging equipment, interphone		
Control Hierarchy	Single stage	Two stages	Three stages
Main Processor	8-bit	8-bit	16-bit
Programming Language	Assembler	Assembler	Assembler, C language
Speech Path Capacity	Trunk link : 4 Intercom : 2	Single-stage time-division, 256-ch non-blocking	
Extension System	Passive bus system		

図３　実験使用データの一例
(OPEN型、罫線が一部除去された英文の表)