節点判別法による表構造理解

1D－10

田中通　吉川大弘　鶴岡信治（三重大学　工学部）

1.はじめに

表構造に関しては，その表の枠の階層的な記述法がよく知られている[1]．本稿では節点行列[2] [3]という，表構造を表現するための新しい概念を提案する．節点行列を用いることで，罫線の一部がない複雑な表（incomplete table）も，罫線が完全な表（complete table）と，同レベルに扱うことができる．また本稿では，節点行列を得るための前処理として，文字消去処理，節点行列抽出処理の計算方法，および，節点行列を応用したHTML文書の作成法について述べる．

2.節点判別法

表の節点は10種類の要素に分類される．各節点の相対的位置を保存し，節点名を行列の要素とすることで，表を行列の形で表現する．これを節点行列と呼ぶ．表1はその10種の節点と，それらが表の要素領域中で，どの部分になり得るのかを示している．

初めに，文字消去処理として，独自に設計された各種ディジタルフィルタを適用し，原画像から表の罫線枠だけを抽出する[2]．それぞれのフィルタは，文字の消去，罫線の両端での伸縮による表の欠損部分の補正等の役割を持つ．これらのフィルタを2値化された原画像（図1）に繰り返し適用することで，文字情報は次第に消えていく．最終的な文字消去処理の結果（罫線図形）を図2に示す．

罫線図形（図2）について，黒画素の縦横方向の周

表1 節点と要素領域中でのなり得る箇所

節点名	左上	右上	左下	右下
n_0 ： 非節点				
n_1 ： ┌	○			
n_2 ： ┬	○	○		
n_3 ： ┐		○		
n_4 ： ├	○		○	
n_5 ： ┼	○	○	○	○
n_6 ： ┤		○		○
n_7 ： └			○	
n_8 ： ┴			○	○
n_9 ： ┘				○

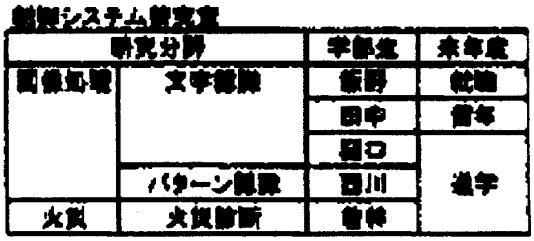


図1 原画像の2値画像

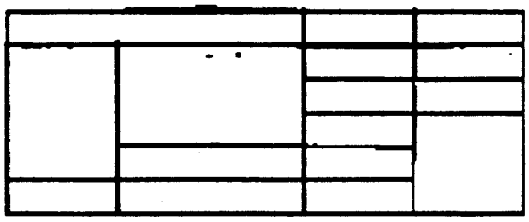
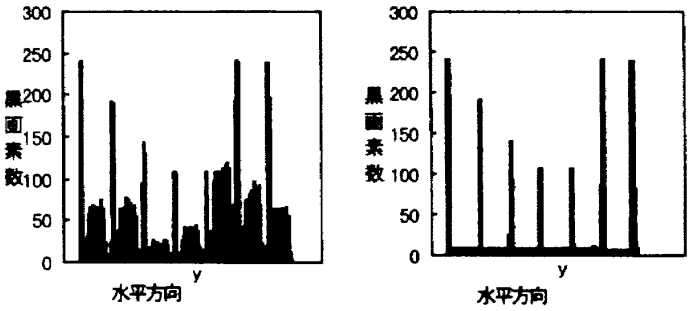


図2 文字消去処理の結果



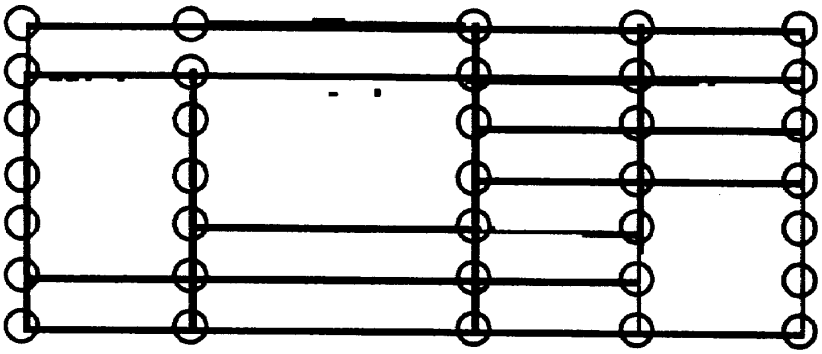
(a)図1の周辺分布　(b)図2の周辺分布

図3 文字消去処理の効果

Table Form Document Understanding Using Node Classification Method

Toru Tanaka, Tomohiro Yoshikawa and Shinji Tsuruoka

Dept. of Electrical and Electronic Engineering, Faculty of Engineering, Mie University



$$A = \begin{pmatrix} n_1 & n_0 & n_2 & n_2 & n_3 \\ n_4 & n_2 & n_5 & n_5 & n_6 \\ n_0 & n_0 & n_4 & n_5 & n_6 \\ n_0 & n_0 & n_4 & n_5 & n_6 \\ n_0 & n_4 & n_5 & n_6 & n_0 \\ n_4 & n_5 & n_5 & n_6 & n_0 \\ n_7 & n_8 & n_8 & n_8 & n_9 \end{pmatrix} \quad (1)$$

図 4 節点位置推定箇所

辺分布を求めることで，文字部分の画素の影響を受けずに，縦横の罫線の位置が推定できる（図 3）．それらの交点位置（図 4 中の丸印が示す箇所）から，節点の位置を特定する．そして，それらの節点の近傍状況より，表 1 に従って，節点名をつける．それらの相対的位置関係を保存し，(1)式のような節点行列を求める．

3.HTML 文書作成法

表の要素領域（罫線で囲まれた領域）の大きさは，節点行列の並びから求めることができ，それらをもとに，HTML 文書で表構造を表すことができる．

まず，領域の左上になる節点行列の要素（表 1 の n_1,n_2,n_4,n_5）に注目し，その要素から右方向にある領域の右上になる要素（n_2,n_3,n_5,n_6）までの距離を調べ，その表の要素領域の幅とする．そして，再び注目点を元に戻し，次はその要素から下方向にある領域の左下になる要素（n_4,n_5,n_7,n_8）までの距離を調べ，その表の要素領域の高さとする．

式(1)から作成された HTML 文書を，WWW ブラウザに通した出力を図 5 に示す．また，罫線で囲まれた部分画像を文字認識ソフトウェアの入力画像とすることで，表中の文字情報を含めた HTML 文書を得ることができると考えられる．

4.まとめ

本稿では，表構造における新しい概念である節点行列を提案した．節点行列を得るための前処理として，文字消去処理と節点分類処理を行い，実験によりその動作確認を行い，15 画像の表についてほとんどの表

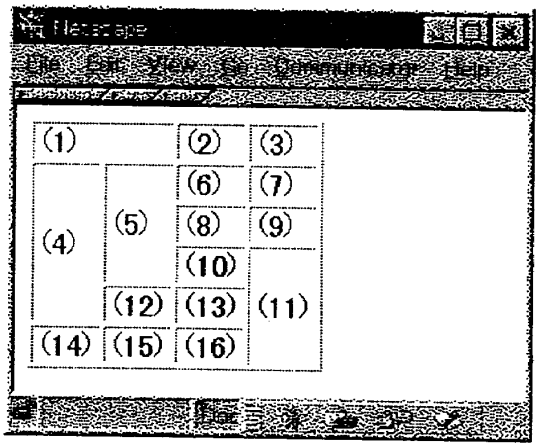


図 5 WWW ブラウザ出力

で良好な結果が得られた．また，得られた節点行列から HTML 文書の作成法を示した．

参考文献

[1] 駱琴,渡邊豊英,杉江昇:"多種帳票文書の構造認識", 電子情報通信学会論文誌(D-II),Vol.J76-D-II,No.10,pp.2165-2176(1993 年 10 月)

[2] 田中通,鶴岡信治,陳新開,石田宗秋:"表形式文書を対象とした節点判別法による HTML ファイルの自動生成",信学技報,Vol.97,No.501,pp.43-50(1998 年 1 月)

[3] T. Tanaka and S. Tsuruoka, "Table Form Document Understanding Using Node Classification Method and HTML Document Generation," Proc. of DAS'98,(査読中)