グラフィクスとCAD 48－5
（1990. 12. 14）

グラフマッチングによる文書画像のストローク抽出法

道庭賢一　佐藤　誠
東京工業大学精密工学研究所

あらまし　情報化社会の高度化に伴い，大量の文書画像データを効率よく処理することがますます重要な問題となっている．伝達，理解，管理など文書画像の処理は従来別々に扱われてきたが，これらは本来統合して処理されるべきものである．このためには文書画像を階層的に表現することが必要であり，ここでは階層表現に適した新しい2値画像の表現法，最小被覆ラン表現（ＭＣＲ表現）を提案する．これは2値画像を最小個数の縦ランと横ランで表現する方法である．本研究では，ＭＣＲ表現を求める問題が2部グラフの最大マッチング問題に帰着することを示すとともに，効率のよいアルゴリズムを明らかにした．

A Stroke Extraction Method of Document Images Based on Matching of Bipartite Graph

Kenichi Douniwa　　Makoto Sato

Research Laboratory of Precision Machinery and Electronics
TOKYO INSTITUTE OF TECHNOLOGY

Abstract　With the advance of information-oriented society, it is necessary to treat a great deal of document images efficiently. Document image processing, such as coding, understanding, and managing, have been treated separately. But the integration of documen image processing, based on hierarchical structures of document images, is desirable. In this paper we propose a new expression of binary images, called Minimum Covering Run Expression. The expression is suitable for hierarchical description of binary images. It is shown that the expression method is closely related to the maximum matchimg problem of bipartite graph. The efficient algorithm for the expression method is also demonstrated.

1　まえがき

情報化社会の発展にともなって膨大な量の文書画像が取り扱われるようになり，この文書画像データを効率よく処理することが重要な課題となっている．

文書画像の基本的な処理としては，ファクシミリなどによる伝達，文字領域・図形領域・画像領域への領域分割や文字認識等の文書理解，データの蓄積や検索によるデータベース管理などが考えられる．これらの伝達，理解，管理などの処理は従来独立して処理されてきたが，本来統合して処理されるべきものである．例えば，ＣＣＩＴＴ（国際電信電話諮問委員会）においてすでに標準化されているＭＨ法，ＭＲ法は，ファクシミリにおける伝達のための効率の良い符号化であるが，理解，管理に関しては必ずしも十分に考慮されていない．

文書画像の構造は図１のように”ページ”，”フレーム”，”ブロック”，”ストローク”，”ラン”，”画素”という構成要素を用いて階層的に表現できる．文字，図形，イメージからなる文書画像を，このような階層構造により表現するための標準化の検討がＣＣＩＴＴとＩＳＯにおいて行なわれており，ＯＤＡ(Open/Office Document Architecture)[1][2]と呼ばれている．また階層構造を利用した文書処理システムの検討も盛んに行われている[3][4]．この階層構造に基づいた表現が実現すれば，伝達，管理のための符号化と，理解のための解析を統合して処理することができる．

文書画像の処理の統合を考えた場合，入力される文書画像から自動的にその階層的表現を得る問題が重要である．２値画像の基本的な表現法としては，ビットマップ表現，横ランを要素とした表現などがある．しかし，これらは階層構造を表現することは考慮されていない．ここでは階層表現に適した新しい２値画像の表現法として，最小被覆ラン表現(Minimum Covering Run Expression)を提案する．以下ではこれをＭＣＲ表現とも呼ぶ．ＭＣＲ表現は縦ランと横ランを基本要素として，２値画像を最小個数のランで被覆することにより表現する方法である．２値画像の縦ランと横ランの集合に，２部グラフ(Bipartite Graph)の構造を導入することにより，ＭＣＲ表現を求める問題は，２部グラフの最大マッチング問題に帰着できる．

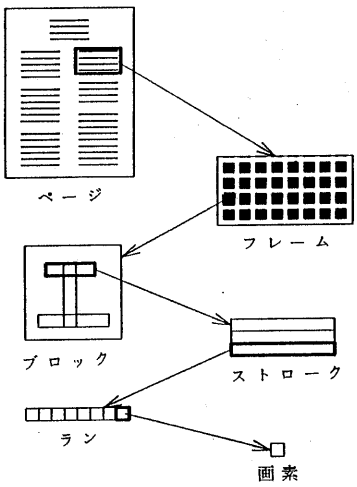


図１　文書画像の階層表現

ＭＣＲ表現の縦ラン被覆部分，横ラン被覆部分のセグメントは，それぞれ縦ストローク，横ストロークに良く対応するため，階層的表現に適し，文書画像の構造解析にも有効である．

２章で，２部グラフとその最大マッチングに関する準備を行なう．３章で２値画像のＭＣＲ表現を定義する．４章でＭＣＲ表現のための解析アルゴリズムを明らかにし，５章で実験結果を示す．

2　２部グラフと最大マッチング

2．1　２部グラフ

頂点の集合をＶ（Ｇ），辺の集合をＥ（Ｇ）とする．互いに素な集合ＸとＹに対して

$V(G)=X \cup Y$

$E(G)=\{<x_i, y_j> \mid x_i \in X, y_j \in Y\}$

となるグラフをＸとＹを部集合とする２部グラフという．ただし$<x_i, y_j>$は頂点x_i，y_jの間の辺を表す．２部グラフは隣接行列により表現することができる．すなわち２部グラフＧの隣接行列

$F(X,Y)=[f_{ij}]$

を次のように定義する．$F(X,Y)$は各要素f_{ij}を

$$f_{ij}=\begin{cases}1: <x_i, y_j> \in E(G)\\ 0: \text{そうでないもの}\end{cases}$$

とする｜Ｘ｜×｜Ｙ｜行列である．ただし

| X |, | Y | はそれぞれ集合X, Yの要素の数である. 2部グラフとその隣接行列の例を図2（a）, （b）に示す.

2. 2 最大マッチングと最小点被覆

辺の集合M⊂Eにおいて, どの相異なる2つの辺$e_i, e_j \in M$も端点を共有しないとき, Mを2部グラフGのマッチングという. つまり行列F(X,Y)の各行, 各列において非零要素をたかだか1個選んだとき, この集合Mがこの2部グラフのマッチングとなる. 2部グラフGのマッチングのうちで, 辺の数が最大のものをGの最大マッチングという. 図2の2部グラフの最大マッチングは, 図2（a）の太い線に対応する辺の集合である. 図2（b）では丸で囲まれた非零要素の集合がそれに対応する. したがって, この場合の最大マッチングMは

$M = \{<x1, y2>, <x2, y1>, <x3, y3>\}$

となる.

次に頂点の集合W⊂Vがどの辺e∈Eの端点をも少なくとも1個含むとき, Wを2部グラフGの点被覆という. つまり行列F(X,Y)のすべての非零要素を被覆するように行および列に直線を引いたとき, この直線に対応する頂点がGの点被覆になる. そして2部グラフGの点被覆のうちで頂点の数が最小のものを最小点被覆という. 図2の2部グラフの最小点被覆は, 図2（a）の黒点に対応する頂点の集合である. 図2（b）では, 行および列に引いた直線に対応する頂点の集合がそれに対応する. したがって最小点被覆Wは

$W = \{x2, x3, y2\}$

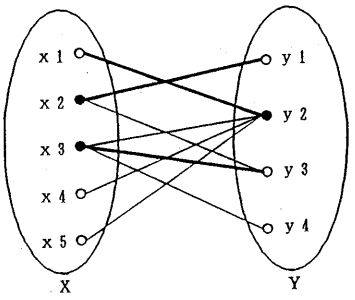
となる.

2部グラフの最大マッチングM, 最小点被覆Wに対して, | M | = | W | が成り立つことがKönigの定理より導かれる[5]. そして2部グラフの最大マッチングを求める問題と最小点被覆を求める問題とは互いに等価である.

2部グラフの最大マッチング問題は, 一般グラフに対する場合よりもはるかに効率的に求めることができる. 例えば Hopcroft-Karp は

$O(\sqrt{n} \cdot (m+n))$

ただしm = | E |, n = | V |



（a）

$F(X,Y) =$

	y1	y2	y3	y4
x1	·	①	·	·
x2	①	·	1	·
x3	·	1	①	1
x4	·	1	·	·
x5	·	1	·	·

（b）

図2 2部グラフと隣接行列

の手続き数で求めるアルゴリズムを提案している[6].

3 2値画像の最小被覆ラン表現

3. 1 最小被覆ラン表現（MCR表現）

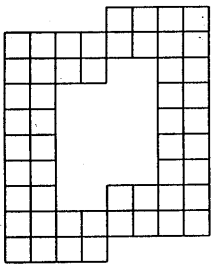
2値画像の横ランを集合Xの要素とし, 縦ランを集合Yの要素とする. 互いに交差し合うランの対に2部グラフの辺を対応させると, 縦ランどうし横ランどうしは交差しないため, 辺は必ず一方の端点をXの集合内に, 他方の端点をYの集合内に持つ. このようにして2値画像に2部グラフを導入することができる.

MCR表現とは, 縦ランと横ランを基本要素として, 最小個数のランで黒画素領域を被覆することにより2値画像を表現する方法である. したがってMCR表現を求める問題は, 2部グラフの最小点被覆問題, あるいは最大マッチング問題に帰着することができる.

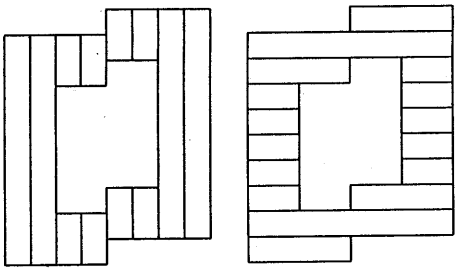
3．2　ＭＣＲ表現とストローク要素

文書画像の多くはストロークを基本要素として構成されている．文書画像中の縦ストロークは横ランで被覆するより，縦ランで被覆するほうがランの数が少なくなる．横ストロークも同様に，横ランで被覆するほうがランの数が少なくなる．

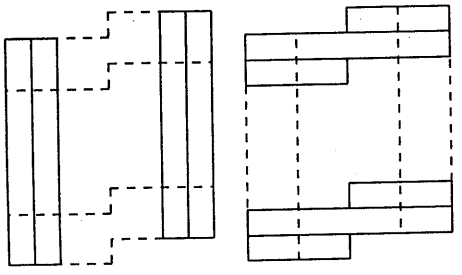
逆に，ＭＣＲ表現の縦ラン被覆部分，横ラン被覆部分のそれぞれの連結要素であるセグメントは，縦ストローク，横ストロークに良く対応していると考えられる．



（ａ）２値画像パターン



（ｂ）２値画像パターンのラン要素



（ｃ）ＭＣＲ表現

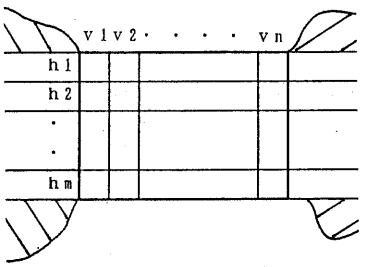
図３

具体例を図３（ａ）に示す．このパターンの縦ランと横ラン要素を図３（ｂ）に示す．このＭＣＲ表現を求め，縦ラン被覆部と横ラン被覆部をそれぞれ図３（ｃ）に示す．縦ラン被覆部，横ラン被覆部の各セグメントがそれぞれ縦ストローク，横ストロークに対応していることが分かる．

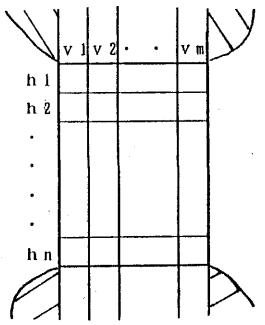
3．3　ＭＣＲ表現のための矩形解析

文書中に大きな表やグラフなどがある場合，これらの部分は黒画素の連結成分の数が一般には大きなものとなる．この部分のＭＣＲ表現を得るには，大規模な２部グラフの最大マッチングを求めることになり，処理時間が問題となる．そのため，ランデータを登録する前処理の段階で，局所的に形状を判断することにより，予めＭＣＲ表現の要素になるラン（被覆ラン）と，ならないラン（非被覆ラン）をある程度確定できると良い．

図４（ａ），（ｂ）のような矩形を含む領域を考えることにする．このとき次の定理が成り立つ．



（ａ）



（ｂ）

図４　最大マッチングの矩形解析

［定理］　図４（ａ）のように上辺と下辺を境界とする縦ｍ，横ｎ（ｍ≦ｎ）の横長の矩形を含む領域について，横ランｈ１，ｈ２，・・・，ｈｍはＭＣＲ表現の被覆ランで，縦ランｖ１，ｖ２，・・・，ｖｎは非被覆ランである．

同様に図４（ｂ）のように，左辺と右辺を境界とする縦ｎ，横ｍ（ｍ≦ｎ）の縦長の矩形を含む領域について，縦ランｖ１，ｖ２，・・・，ｖｍはＭＣＲ表現の被覆ランで，横ランｈ１，ｈ２，・・・，ｈｎは非被覆ランである．

この定理により図４（ａ），（ｂ）のような矩形領域を調べることにより，前処理の段階で多くのランを被覆ラン，非被覆ランに確定できる．そして最大マッチング問題の対象となる領域は，図４の斜線部で示した部分領域である．これらの領域はもとの２値画像に比べてはるかに細分化されているので，処理の高速化が期待できる．この矩形領域の局所的処理を，以後ＭＣＲ表現のための矩形解析と呼ぶ．

4　ＭＣＲ表現の解析アルゴリズム

4．1　ランデータ構造

最大マッチンで必要になる隣接行列を生成するためには，交差するランが効率よく求められるように縦ラン，横ランのデータを管理する必要がある．横ランデータの管理方法は図５（ｂ）のようにする．

横ランデータを管理するために，二つの配列を用意する．配列hran()は，横ランデータの始点座標(sp)，終点座標(ep)，フラグ(fg)，および同一走査線状の次のランデータの保存されている番地を表すポインタ(np)から構成される．フラグ(fg)は，被覆ラン，非被覆ラン，あるいはいずれにも確定していない未処理ランなどのランの属性を表現する．htop()は，各水平走査線に属する最初の横ランデータのポインタを設定する．この横ランのデータ構造により，任意の水平走査線上のランデータを効率よく取り出すことができる．

縦ランのデータ構造も，全く同様にして２つの配列vran()とvtop()により表現される．

このようなランデータ構造を用いることにより，交差するランは次のようにして求めることができる．

垂直走査線ｙ上で，始点座標がxs，終点座標がxeの縦ランと交差する横ランは

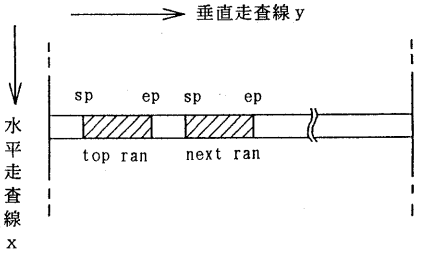
$$xs \leqq x \leqq xe$$

を満たす水平走査線ｘに対して以下の処理を繰り返すことによって求められる．水平走査線ｘ上の最初の横ランデータのポインタhtop(x)により，水平走査線ｘ上の最初の横ランデータを参照する．この横ランデータの始点座標sp，終点座標epが

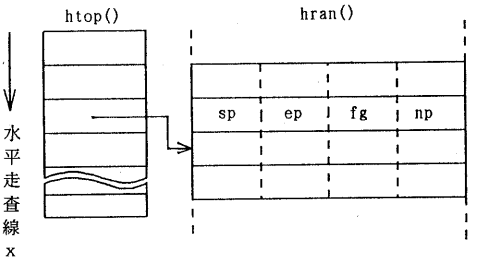
$$sp \leqq y \leqq ep$$

の条件を満たしているとき，この横ランが元の縦ランと交差するランである．そうでないときは，ポインタ(np)により次のランを調べる．この操作を交差するランが求められるまで続ける．交差する横ランは１つの水平走査線に１つだけ存在するので，交差するランが見つかったら，次の水平走査線の処理に移る．

横ランと交差する縦ランも，同様にして効率よく求めることができる．



（ａ）　２値画像の水平走査線データ



（ｂ）　横ランデータ構造

図５

4．2　解析アルゴリズム

ＭＣＲ表現を求める解析アルゴリズムのフローチャートを図６に示し，以下で説明する．

① ２値画像データを水平走査線ごとに読み込みながら縦ランと横ランのデータを登録し，同時に矩形解析も行う．この処理の詳しい説明は以下で示す．

② 未処理ランがなくなるまで，③と④の処理を繰り返す．

③ ②で検索された未処理ランと交差する未処理ランを求め，さらにそのランと交差する未処理ランを求めるというように，再帰的に交差する未処理ランを求めながら，隣接行列を作成する．

④ ③で作成した隣接行列に対して最大マッチング問題を解くことにより，最小被覆ランを求める．本研究ではHopcroft-Karpのアルゴリズムを用いて最大マッチングを解く(6)．

①のアルゴリズムについて詳しく説明する．

ランデータの登録のために，現時点での水平走査線データと，１つ前の水平走査線データのためのラインバッファnewline()とoldline()を用意する．

横ランデータの登録は，ラインバッファnewline()の黒画素成分を調べることにより行う．

縦ランデータの登録は，２つのラインバッファを比較することにより行う．図７（ａ）に示すように，２つのラインバッファの比較により４つの状態変化

Ⅰ　白画素から白画素
　・縦ランが存在していない
Ⅱ　白画素から黒画素
　・縦ランが生成
Ⅲ　黒画素から黒画素
　・縦ランが継続
Ⅳ　黒画素から白画素
　・縦ランが終了

が考えられる．状態変化Ⅱでは，新しい縦ランを登録し始点座標を与える．状態変化Ⅳでは，すでに登録してある縦ランの終点座標を与える．

図４（ａ）の形状の矩形解析について述べる．このような形状が存在するのは，長さｍの縦ランがｍ個以上横方向に連続して終了した場合である．このことから状態変化Ⅳの縦ラン終了時に，縦ランの長さと繰り返し回数を調べることにより，図４（ａ）の矩形領域を判断することができる．

次に図４（ｂ）の矩形解析について述べる．この矩形領域が存在するのは，白画素から黒画素への境界と黒画素から白画素への境界のペア

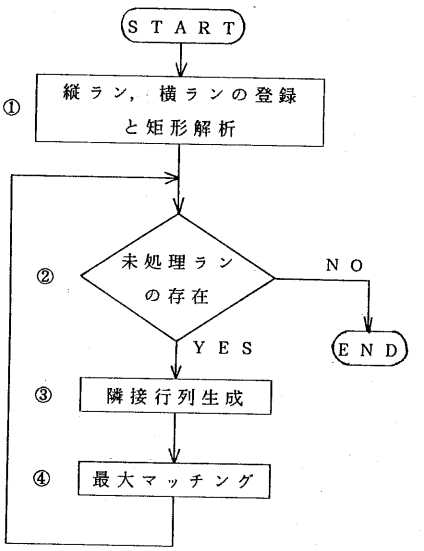
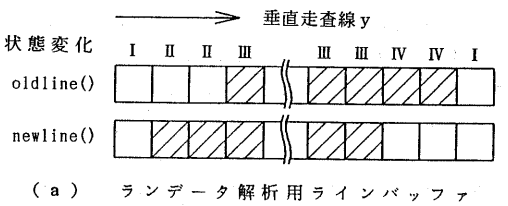
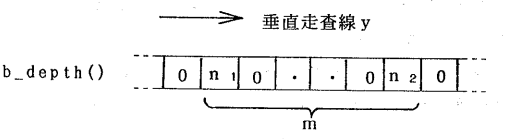


図６　ＭＣＲ表現の解析アルゴリズム



（ａ）　ランデータ解析用ラインバッファ



（ｂ）　図４（ｂ）の矩形解析

図７　解析用データ

が幅mをもち，m個以上縦方向へ続いた場合である．このため，境界の縦方向の連続数を示す配列b_depth()を用意する．b_depth()には白画素から黒画素への境界の縦方向の連続数，あるいは黒画素から白画素への境界の縦方向の連続数を設定し，境界でないときには0にする．b_depth()は水平走査線を読み込むごとに，常に更新する．図7（b）のように，b_depth()の境界の対n_1とn_2，その間の幅mに対して，

$$m \leq \min\{n_1, n_2\}$$

が成り立つとき，この領域が図4（b）の矩形領域にあたる．

図4（a），（b）の矩形領域が検出された場合，該当するランを被覆ランか，非被覆ランに確定する．

5 実験結果

ＣＣＩＴＴ標準原稿のＭＣＲ表現を求め，その実験結果を示す．

ＣＣＩＴＴ標準原稿は，大きさが縦２３７６，横１６８０の２値画像データである．その例を図8（a）に示す．この例についてＭＣＲ表現を求め，その縦ラン被覆部，横ラン被覆部をそれぞれ図8（b），（c）に示す．この実験結果により文書画像の表を構成する縦ストローク，横ストロークが，それぞれ縦ラン被覆部，横ラン被覆部のセグメントとして抽出されていることが分かる．文字に対してもストロークのはっきりしている部分は，ある程度ストロークと縦ラン被覆部，横ラン被覆部のセグメントとの対応がとれていることが分かる．

次にＭＣＲ表現の応用例として，文書画像から表を抽出する問題を考えてみる．表を構成している各ストロークの長さは，文字の大きさに対して十分に長い．このことから適当な閾値を定めて，縦ラン被覆部，横ラン被覆部の各セグメントの形状を分析することにより表部分の抽出を行うことができる．図8（a）の文書画像から表を抽出した結果を図8（d）に示す．

6 まとめ

本研究では，２値画像の新しい表現法としてＭＣＲ表現を提案した．この表現法は，２値画像を最小個数の縦ランと横ランで被覆することにより画像を表現する方法である．２値画像からＭＣＲ表現を求める問題は，縦ランと横ランを要素とする２部グラフの最大マッチング問題に帰着することを示した．さらにＭＣＲ表現を求めるための，効率のよい解析アルゴリズムを明らかにした．また，文書画像に対するＭＣＲ表現の実験結果を示し，文書画像解析への可能性を示した．

ＭＣＲ表現はできるだけ少ないランで２値画像を表現するという意味で情報圧縮を行っており，効率のよい文書画像の符号化の可能性をもっている．しかし，そのためには本手法に適合した符号化法を考える必要があり，今後の課題である．また，具体的な文書画像の解析問題に本表現法を適用して，その有効性を確かめることも課題として残されている．

謝辞 有益な御助言をいただいた東京工業大学精密工学研究所河原田弘教授に深謝いたします．

文献

（１）"Open Document Architecture (ODA) and Interchange Format", CCITT, Rec. T.410 シリーズ(1988)

（２）"Office Document Architecture (ODA) and Interchange Format", ISO 8613 (1988)

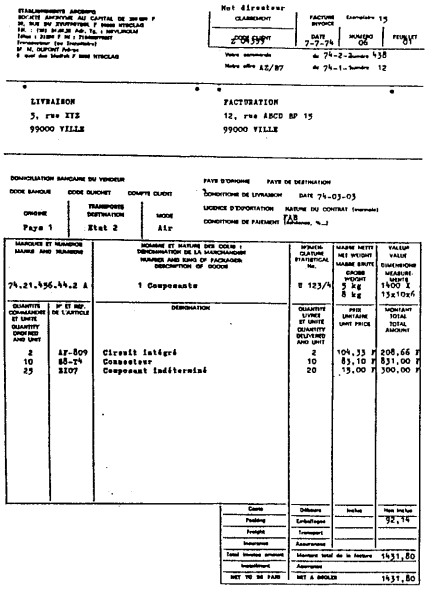
（３）山田，藤長，遠藤，蓮池：”ミクスモード文書作成システム”，画電学誌，15,4 pp.274-282 (1986)

（４）山田，宮里，蓮池：”マルチメディア文書構造処理システム”，画電学誌，19,5 pp.286-295 (1990)

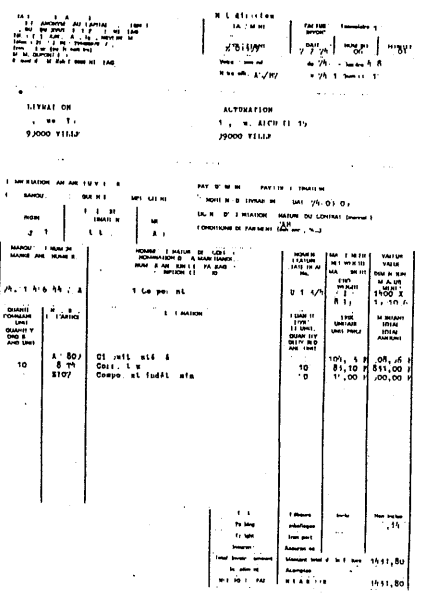
（５）König, D:"Graphen und Matrizen", Mat. Fiz. Lapok, 38, pp.116-119. (1931)

（６）Hopcroft, J.E. and Karp, R.M:"An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs", SIAM, J. Computer, 2,4, pp.225-231 (1973)

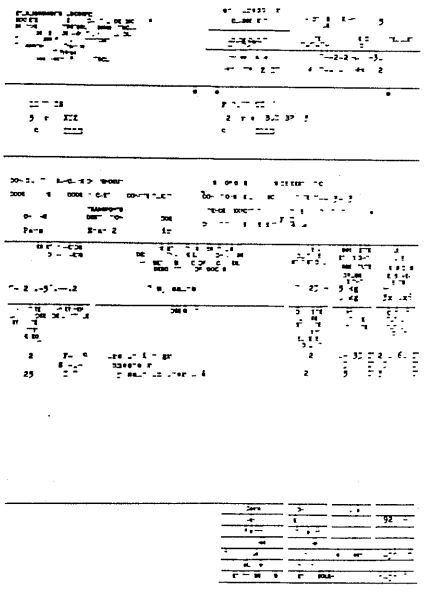
（７）嶋，柏岡，東野：”ランに対する座標演算に基づく２値画像の高速化回転のための一手法”，信学論(D)，J71-D，7，pp. 1296-1305 (1988)



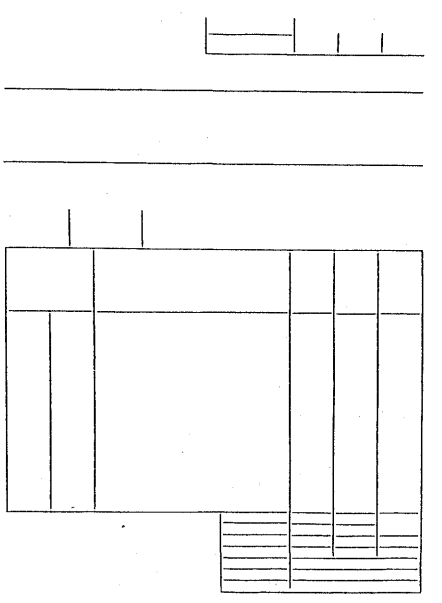
（ａ）原画像（ＣＣＩＴＴ標準原稿ＮＯ３）



（ｂ）縦ラン被覆部



（ｃ）横ラン被覆部



（ｄ）表の部分抽出結果

図８　文書画像のＭＣＲ表現