4E-4 連結成分特徴に基づく文字分離抽出方式

樋野 匡利、 福田 浩至、 町田 哲夫
(日立製作所システム開発研究所)

1．まえがき

印刷文書の自動入力において重要な機能として、
(1) 文字、表、図、画像の各部分の分離、
(2) 文字部の構造情報抽出と文字認識、
(3) 表内の文字の分離、認識、
等がある。

各機能について多くの報告があり、我々も上記(1)、(2)に関して、黒画素連結成分の外接矩形に着目した方法を提案した [1] 。本稿では、この方法を拡張し、(3)の処理と文字切り出しに適用できるようにした内容について報告する。

表内文字の分離では、黒画素連結成分（以下、連結成分と略す）の抽出時に、線と文字のラン長の差に着目して接触部の分離を行ないながら、文字候補要素のみを抽出する。抽出された文字候補要素に対して、連結成分の外接矩形情報を基に、文字列の抽出、文字切り出しを行なう。

2．文字、線の分離、抽出

文字と図形（表も含む）中の線の分離方式については、多くの報告 [2] etc があるが、本稿では、分離する対象を表、アンダーライン等の縦横線に限定し、連結成分の抽出時にラン長の差に着目して処理する方式を提案する。連結成分情報は、文字列の抽出や文字切り出し等の処理に有効であるが、文字と線が接触した場合、その文字と線は同一の連結成分となり、分離できない点が問題となる。多くの場合、連結成分の抽出を行なう前、または、後で、他の特徴に基づいて分離を行なう。これに対して、処理をできる限り簡単化し、以後の処理と整合性を良くするため、連結成分の抽出時に、接触の切り出しを行ないながら、文字と線の分離をする方式をとった。画像データを走査するだけで、連結成分を抽出しながら、文字、線の分離を行なうことができる。

連結成分の抽出は、画像を走査し、各ラインのランに着目し、隣接ライン間でのランとの接続を調べることにより行なう。このとき、文字と線の接触の切り出しを行なうため、

(a) 推定文字幅の上限 w_max よりも長いランは、文字の構成要素ではない、と考え、隣接ラインのランとは、非連結であるとして処理する。
(b) (a)の条件で分離した連結成分に対して、長さの類似するランが m 個以上続かない場合、そのランは線の一部であると考えて、文字候補成分とはしない。この条件は、線のエッジ部分の凹凸によるノイズを除去するためである。
(c) 抽出された連結成分のうち、高さが推定文字高さの上限 h_max 以上のものは、縦線候補とする。
(d) 逆方向の走査に対しても、同様の処理を行ない、二つの結果を統合して、文字候補要素の抽出を行なう。

図1に、表中のアンダーラインと文字が接触している例の処理結果を示す。接触部の分離位置が問題となるが、この例では、よい位置で分離できたと思われる。

研究会	日	時
ヒューマンインタフェース	5月15日(月)	13：00
情報システム	5月15日(月)	13：30
マイクロコンピュータとワークステーション	5月16日(火)	13：30
データベース・システム	5月18日(木)	10：30

研究会	日	時
ヒューマンインタフェース	5月15日(月)	13：00
情報システム	5月15日(月)	13：30
マイクロコンピュータとワークステーション	5月16日(火)	13：30
データベース・システム	5月18日(木)	10：30

図 1 文字、線の分離結果

3．文字列の抽出

文字候補として抽出された連結成分の外接矩形を統合して、文字列を抽出する。隣接する矩形の大きさ、距離、位置関係等に基づいて行なうが、この方法については、既に、報告した[1]。

Character Extraction Method based on the Feature of Connected Components
Masatoshi HINO, Kouji FUKUDA, Tetsuo MACHIDA
HITACHI, Ltd.

4．文字の切り出し

文字の切り出し方式には、文字の大きさを仮定しておこなうもの、ピッチに基づくもの等、多くの方法が提案されている [3]、[4] etc 。

ここでは、文字列を構成する文字の連結成分の外接矩形（以後、矩形と略す）の特徴を用いて、文字を切り出す。処理対象は、全角、半角が混在する文字列で、プロポーショナルピッチとする。ただし、英単語は含まない。仮定する条件は、全角の文字は縦横比が1に近い、ピッチは局所的には安定している、こととする。

切り出しの基本的な考え方を述べる。1) 左右の矩形と統合すると、全角の縦横比条件を満たさなくなり、それ単独で全角と判断される(安定な)矩形をベースとする。2) 効率を考え、矩形情報のみで、可能な限り全角分離文字を統合する。3) 矩形情報では、全角分離文字か、半角の並びか判定できないものに対して、文字認識による切り出しを行なう。

以下、処理を説明する。各文字列中の矩形に対して、

(a) 上下方向で重なる矩形を統合する。

(b) 文字列の高さをもとに、各矩形の高さを評価し、文字の高さ h_ave、幅 w_ave を推定する。

(c) 推定文字幅と縦横比を基準として、安定な全角矩形を抽出する。

(d) それ以外の矩形を、文字高さを基準に半角と「その他」とに分類する。「その他」と推定された矩形は、その左右の矩形と統合し、全て、全角、半角の矩形にする。

このときの統合条件は、

i）統合結果の幅が文字幅の上限を超えない、

ii）左右どちらとも統合可能な時、統合結果の幅の小さいほうの統合を選択する。

この条件を満たして統合できないものは、ノイズとして除去する。

(e) 前後の矩形とのピッチ pr、pl に着目し、半角文字幅より小さい矩形は、前後の矩形と統合する。その条件は、図2に示すように、どちらか一方のみが半角文字幅より小さい場合は、その矩形と、両方とも小さい場合は、矩形間距離 dr、dl 統合結果の幅 wr、wl を求め。

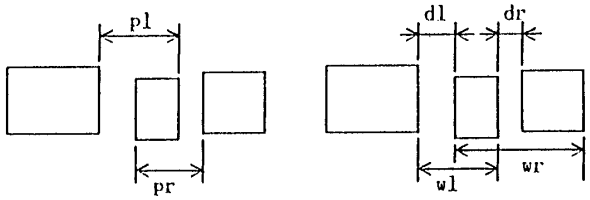
- ((dl>dr)&(wl>wr)) or (dl>dr)&(wl>wr)):
  左の矩形と統合する。
- ((dl<dr)&(wl<wr)) or (dl<dr)&(wl<wr)):
  右の矩形と統合する。
- その他の場合:
  l_eval = min(|wl-h_ave|,|wl-h_ave/2|)
  r_eval = min(|wr-h_ave|,|wr-h_ave/2|) として
  l_eval > r_eval :左の矩形と統合する
  l_eval < r_eval :右の矩形と統合する

(f) 半角の並びに対して、全角の分離文字のある可能性も含めて、全てのケースを求め、各ケースで文字認識を行ない、全体としての認識結果の良いものを、正しい文字切り出し結果とする。

図3に、文字切り出し結果の例を示す。この例では、全角の分離文字と半角文字の混在する文字列に対して、正しく切り出すことができた。

5．むすび

表やアンダーラインを含む印刷文書を対象に、連結成分の抽出時に、接触の切り出しを行ないながら、文字と線を分離、抽出する方式を提案した。また、全角半角の混在するプロポーショナルピッチの文字列に対して、連結成分の外接矩形の性質を用いて、切り出しを行なう方式を提案した。これにより、印刷文書の自動入力を行なう際の処理対象を広げることができると考える。



(a) 前後のピッチ　　(b) 矩形間距離と文字幅

図 2　統合条件

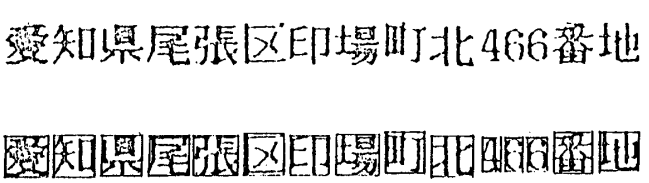


図 3　文字切り出しの結果

参考文献

[1] 樋野 他 ："マルチメディア処理によるオフィスワークステーション - 文書構造の分離抽出方式 -"、情処第32回全大、3K-2

[2] 岩城 他 ："近接線密度法による文字、図形切り分け処理の検討"、信学研資PRL81-81 pp.109-116

[3] 辻 他 ："分散最小基準に基づく適応型文字分離方式"、信学論Vol.J68-D No.8 pp.1497-1504

[4] 村瀬 ："手書き文字列認識における文字切り出しの個人適応化"、信学論Vol.J72-D-II No.1 pp.132-139