

名刺OCRにおける領域抽出

6W-4

中村昌弘 , 足立修
(株)リコー 中央研究所

1. はじめに

漢字認識技術の発達に伴い、漢字OCRの実用化が進んでいる。他方、名刺をデータベース化して管理運用するニーズが高まり、その入力手段として漢字OCRが期待されている。(1)(2)

名刺は会社名、氏名、住所、電話番号等いくつかの項目から構成されており、データベース化するには各項目の属性を認識することが不可欠で、そのためには項目ごとに領域を抽出する必要がある。また各項目は文字の大きさ、ピッチ等が異なるため、項目別に領域を抽出することによりそれぞれの項目に最適な処理を行うことができ、システムとしての性能の向上が期待できる。

本稿では画像の大きさと、相対的な位置関係を使い項目認識の前提となる領域を抽出する手法を報告する。

2. 基本原理

名刺が一行一項目であれば、これまでの単純に射影をとる方法で領域抽出が可能であるが、通常の名刺は図1のように二項目以上の射影が重なっている例が多い。

そこでいったん名刺画像を小ブロックに分解し、名刺画像に関するいくつかのルールを適用しながら小ブロックを統合していき、統合した結果として領域抽出となるような手法を考案した。小ブロックの統合に適用するルールを下記に示す。

- ① 同じ属性の項目の文字はほぼ等しい大きさである。
- ② 同じ属性の項目は同一行にある。
- ③ 文字はその形がほぼ正方形である。

3. アルゴリズムの概要

領域抽出は図2のフローに従って行われる。

3. 1 小ブロックの抽出

名刺の長手方向に射影をとり行を切り出す。(図3-a) 次に各行毎に行の垂直方向に射影をとり、ブロックを切り出す(図3-b)。さらに各ブロック毎に、

行と水平方向垂直方向に射影をとり小ブロックを抽出する。抽出した小ブロックには抽出順に番号をつける(図3-c, d)。

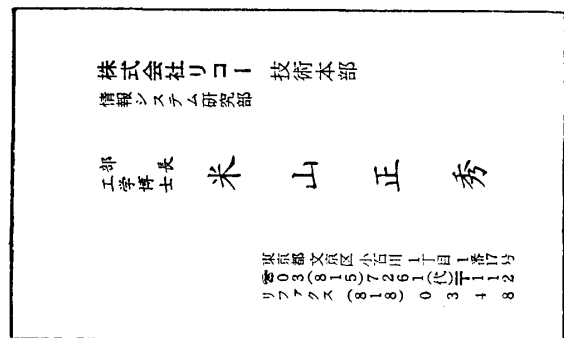


図1 項目が重なる名刺の例

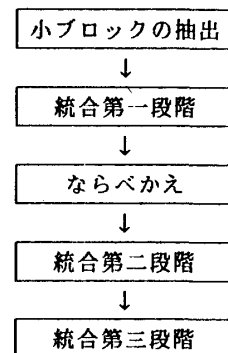


図2 領域抽出の流れ



図3 小ブロックの抽出

3. 2 統合第一段階

3. 2. 1 文字構成ブロックの統合

漢字にはそれぞれ一文字がいくつかの小ブロックに分解されるものも多く、前後の文字と統合する前に一文字の構成を調べる必要がある。そこで基本原理のルール③を適用し次の三項目により複数のブロックで一文字が構成されているか判定し、必要ならまず文字単位に統合する(図4)。

- 1) ブロックの高さと幅の比
- 2) ブロックの高さと行の高さ
- 3) ブロックの幅

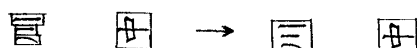


図4 文字構成ブロックの統合

3. 2. 2 ブロックの統合

3. 2. 1で文字単位となったブロックに対し基本原理のルール①②を適用し、抽出の番号順に連続する二つのブロックの高さ、お互いの位置関係の情報により一つブロックとして統合するかを判定する。

3. 3 ブロックのならべかえ

統合第一段階が終了するとほとんどのブロックは大きさ別に統合されるが、射影が重なる項目には図5-aのように同一行が連続して統合されないことがある。図5-bは統合第一段終了時にブロックに付加された番号である。そこで統合されたブロックを次のルール(図5-c参照)によりならべかえる(図5-d)。

- 1) 注目するブロックのトップ位置が次のブロックのボトム位置より下にある。
- 2) 注目するブロックが次のブロックより左にある。
(左から右に走査しているものとする)

3. 4 統合第二段階

ならべかえたブロックに対し再度番号順に連続する二つのブロックに対し統合処理を行う。統合の判定ルールは統合第一段階と同じである。ここで図5は図6のように統合される。

3. 5 統合第三段階

図7のような途中で極端に小さいブロックが存在する画像の場合、統合第二段階までの処理では図7-aのようになる。これに対応するため統合第三段階では、統合を妨げる原因となるブロックの前後のブロックに対し、ブロックの高さとお互いの位置関係の情報から統合するか判定する。

この一連の処理により小さく分解されたブロックがその大きさ別にいくつかのブロックに統合される。これらの統合されたブロックが名刺画像の各項目領域になる。

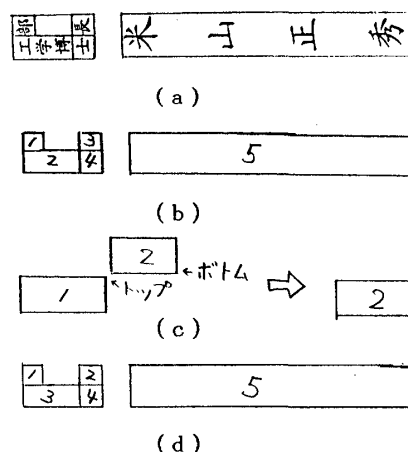


図5 並べ替え

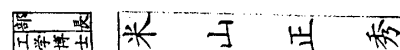


図6 統合第二段階終了後

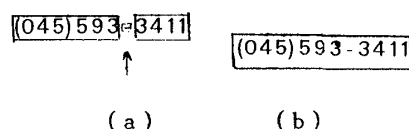


図7 統合第三段階の例

4. 実験結果

縦書き名刺、横書き名刺各100枚について抽出実験を行った。表1がその結果である。

横書き名刺の方が縦書き名刺に比べ領域抽出率が悪いのは、同一行に異種項目(例えば住所と電話番号)が混在するものがあり、基本原理のルール②に適用しないためである。

領域抽出率 = 正抽出領域数 / 全項目数

表1 領域抽出率

型	領域抽出率
縦書き	98.0%
横書き	91.9%

5. まとめ

画像を小ブロックに分解し、ブロックの大きさと位置関係の情報を使った簡単なルールを適用してブロックの統合を行うことで名刺項目の領域抽出が実現できた。

<参考文献>

- (1) 山田他: "文書画像の構造理解における文字列の抽出について", 昭63春信学全大SD-7-5
- (2) 黄瀬他: "名刺画像認識における項目仮説生成", PRU87-88