ＰＣ上で動く印刷英数字ＯＣＲ

5P-3

中村洋治，除村健俊，豊川和治，北山友
(日本アイ・ビー・エム株式会社　大和研究所　画像システム開発)
高橋弘晏
(日本アイ・ビー・エム株式会社　サイエンス・インスティチュート)

１．まえがき

本論文では、特定化されていないフォントが混合されている文書を読むことができ、かつ処理が簡単な文字認識アルゴリズムについて述べる。このアルゴリズムは、読み込まれた文字から特徴を抽出し、階層的に２段階の大分類を行なった後、識別を行なう。さらに、類似文字の１対１の比較によるチェックを行なって認識率を高めている。

２．認識アルゴリズム

文字の認識は図１に示すような手順で行なわれる。

（１）イメージの入力：　文字を含むイメージ情報が入力され、記憶装置上にたくわえられる。

（２）文字の切り出し：　一文字ごとにそのイメージが切り出され、32×32画素の領域に格納される。

（３）特徴抽出１　：　文字の各画素ごとに2×2のマスクをかけ、その中の白黒のパターンを調べることにより、局所輪郭特徴を抽出する。特徴抽出の様子を図２に示す。抽出された特徴量は図３に示すような領域ごとに集計され、4（領域の分割方向）×6（分割された区画の数）×4（局所輪郭の方向）＝96次元の特徴量とされる。以後、このように分割された領域をルック領域、分割の方向をルック方向と呼ぶ。この特徴は識別のために使用されるが、大分類１のためにはルック領域に関係なく4つの局所輪郭方向別に合計し、4次元の特徴量を求める。

（４）大分類１　：　（３）で求めた入力文字の4次元特徴量について、予め用意してある全てのカテゴリーに対応する標準パターンとの距離を次式で計算し、カテゴリーごとに定められた閾値を越えるカテゴリーを候補からはずす。

$$4\text{次元距離} = \sum_{i=1}^{4} | f_i - g_i |$$

ただし、　f_i：標準パターンの4次元特徴量；　g_i：入力パターンの4次元特徴量；　i：局所輪郭の方向

（５）特徴抽出２　：　図４のように、32×32中30×30の領域で4辺から中心に向かって進み、イメージ領域の半分に達するか、文字にぶつかるまでの距離を計る。4辺をそれぞれ6区画に分割し、その区画内で距離の合計を求めることにより、4×6＝24次元の特徴を抽出できる。

（６）大分類２　：　（５）で求めた入力文字の24次元の特徴量について、大分類１を通過したカテゴリーに対応する標準パターンとの距離を次の式で計算し、カテゴリーごとに定められた閾値を越えるものを候補からはずす。ここで、文字の大きさの変化による影響を打ち消すために、標準パターン、入力パターンの特徴量の平均m_f，m_gを用いている。

$$24\text{次元距離} = \sum_{i=1}^{4}\sum_{j=1}^{6} | (f_{ij} - m_f) - (g_{ij} - m_g) | = \sum_{i=1}^{4}\sum_{j=1}^{6} | f_{ij} - g_{ij} + (m_g - m_f) |$$

ただし、　f_{ij}　：　標準パターンの24次元特徴量　　i　：　辺の番号
　　g_{ij}　：　入力パターンの24次元特徴量　　j　：　区画の番号

$$m_f = (\sum_{i=1}^{4}\sum_{j=1}^{6} f_{ij}) / 24 \qquad m_g = (\sum_{i=1}^{4}\sum_{j=1}^{6} g_{ij}) / 24$$

（$m_f - m_g$は１度だけ計算する。）

（７）　識別　：　次式により、大分類１と２を通過したカテゴリーの標準パターンと入力パターンとの距離を求め、各候補を距離の近い順に並べる。Ｋは24次元距離と96次元距離との間の大きさの違いを補正するための定数である。

$$\text{距離} = 24\text{次元距離} + K \times 96\text{次元距離}$$

96次元距離は次の式で求められる。

$$96\text{次元距離} = \sum_{i=1}^{4}\sum_{j=1}^{4}\sum_{k=1}^{6} w_{ij} | f_{ijk} - g_{ijk} |$$

ただし、　f_{ijk}：標準パターンの96次元特徴量；　i：局所輪郭の方向；　k：ルック領域
　　g_{ijk}：入力パターンの96次元特徴量；　j：ルック方向；　w_{ij}：重み定数

ここで、w_{ij}の値は次の式で決定される。

$$w_{ij} = w_i \times w'_{ij}$$

ただし、　w_i　：　局所輪郭の方向iによる重み　　w'_{ij}　：　iとルック方向jとの間の角度による重み

w_iは局所輪郭のマスク中での長さを考慮したものである。その値は、例えば次のようになる。

w_i　＝　３　（i＝1，3　のとき）、又は　２　（i＝2，4　のとき）

w'_{ij}は文字のずれによる影響を考慮したものである。一般に、ルック領域と輪郭線が垂直な場合に比べ、平行な場合の方が文字のずれが特徴量に及ぼす影響が大きい。そのため、w'_{ij}の値は、局所輪郭方向とルック方向のなす角度が直角に近いほど大きく、平行に近いほど小さくしている。w'_{ij}の値の例を次に示す。

Software OCR for Typed English Characters

Y. Nakamura, T. Yokemura, K. Toyokawa, Y. Kitayama, H. Takahashi
IBM Japan

w_{ij} ＝ 1 （角度：0° のとき）、3 （角度：45° のとき）、又は 4 （角度：90° のとき）

（8） 類似文字識別 ： 一般に、類似している文字は全体のパターンの相違より局所的なパターンの相違により区別しやすいので、局所的特徴を用いて認識率を向上させる。そのため、（7）で並べられた候補の中から次の2つの条件を共に満たすCiについて、局所パターンを用いた詳細チェックを行う。ただし、候補を距離の近い順にC_1、C_2、...とする。

◇ 入力パターンとC_iとの距離 ≦ 入力パターンとC_1との距離×M

◇ i ≦ N

ただし、M、Nは経験的に求めた定数であり、例えば英数字に関しては、1.25と5をそれぞれ使っている。こうして残った候補をあらたにC_1、C_2...C_nとすると、類似文字識別の手順は次のようになる。

【1】2候補C_n、C_{n-1}について、次の(i),(ii)を行う。

(i) 予め定義された類似文字識別用特徴テーブルを見て、24次元と96次元の特徴の中からこの特定の2候補を区別するのに最も適した局所特徴を数個選ぶ。ここでは2個にしている。このテーブルの作成方法については後述する。

(ii) 入力パターンと2候補との距離を次の式によって夫々求め、近い方を勝者とする。ただし、(ii)で選ばれた2つの特徴をi，jとする。

$$距離 = K_i|f_i - g_i| + K_j|f_j - g_j|$$

ただし、f_i，f_j ： 標準パターンの特徴量　　g_i，g_j ： 入力パターンの特徴量

K_i，K_j ： 24次元距離と96次元距離との間の大きさの違いを補正するための重み定数

（1≦i≦24ならばf_iは24次元の第i番目の特徴、25≦i≦120ならば96次元の第（i－24）番目の特徴を表わす。）

【2】C_n，C_{n-1}間の勝者とC_{n-2}とについて、(i),(ii)を行う。

【3】こうして勝ち抜き制で比較をしていき、最後にC_1とその前の勝者とを比較して距離の近かった方を最終結果とする。

この方法では、すでに抽出された特徴のうちからいくつかを選んで使っているため、新たに特徴抽出を行う手間が不要であり、また一時に2つずつの候補を取り出してきて比較することにより、簡単できめ細かいチェックが可能である。

3．類似文字識別のための局所特徴テーブルの作成

類似文字識別において2つの候補を比較するために使われる特徴は、次の式の値を最大及び次大にするように120次元の特徴の中から予め選ばれ、局所特徴テーブルに納められたものである。

$$(M_{ix} - M_{jx})^2/(V_{ix} + V_{jx}) \qquad (1 \leqq x \leqq 120)$$

ただし、M_{ix} ： カテゴリーC_iのx番目の特徴量の平均　　V_{ix} ： カテゴリーC_iのx番目の特徴量の分散

M_{jx} ： カテゴリーC_jのx番目の特徴量の平均　　V_{jx} ： カテゴリーC_jのx番目の特徴量の分散

ここでいう平均、分散とは標準パターン作成時に一つのカテゴリーに対し使用する複数のトレーニング・データから求められるものである。

4．むすび

本論文で述べたアルゴリズムにより、簡単な手法でマルチフォント印刷文字の分類、識別、類似文字のチェックを行なうことができる。実験では、IBM-PC/AT (80286,6MHz)を使い、2回コピーされた文書に対して、99%以上の認識率と約7.5文字／秒の認識速度が得られた。

5．参考文献

（1）高橋：＂細線連結素方向による簡易手書漢字認識＂，信学技報PRL82-8 (1982)

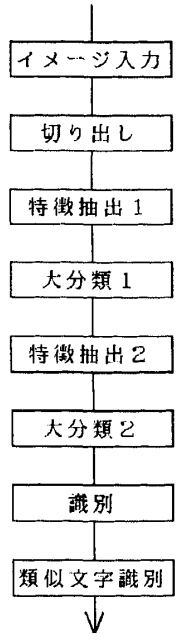

図1 全体の処理の流れ

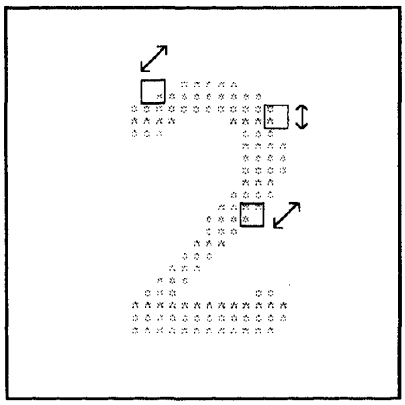

図2 局所輪郭特徴の抽出

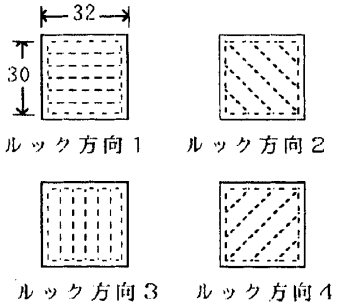

図3 特徴抽出1における領域の分け方

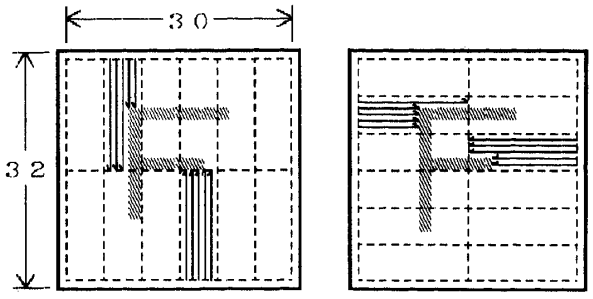

図4 外形特徴の抽出