I-016

自由文読取のための文字切出し方式の開発

Development of Character Segmentation Method for Recognition of Unconstrained Handwriting String

古川 直広† 徳野淳子‡ 池田 尚司†
Naohiro Furukawa Junko Tokuno Hisashi Ikeda

1. はじめに

近年の技術進歩に伴い、住所や氏名など表記が既知の文字列だけでなく、質問文やコメント文など記入者が自由に記入した文字列を認識するニーズが高まっている。

このため、自由に記入された文字列を認識するための文字切出し方式について検討した。利用可能な特徴量として、文字サイズや文字共起情報など種々の情報が考えられる。本報告では、文字パターンの幾何情報と時間情報、文字識別結果に着目し、確率モデルを用いて文字パターン尤度を算出する方式を検討した。その方式の詳細とデジタルペン[1]を用いた評価実験について報告する。

2. 課題とアプローチ

2.1 文字認識と文字切出し

文字認識部の処理は、大きく分けて、

(1) 文字切出し：入力文字行中から文字らしいパターン(文字パターン)を抽出、
(2) 文字識別：各文字パターンを文字識別、
(3) 文字列照合：文字識別結果を意味のある文字列として解釈、

の 3 ステップからなる(図 1)。

文字切出しにおいて、個々の文字パターンの形状情報のみでは切出し位置が一意に決定できない。そのため、考えうる文字パターン候補を作成(プレセグメンテーション)し、これをグラフ辺とする非循環有向グラフ(文字切出しグラフ)で表現する(図 1)。各文字パターンに対し、その尤もらしさを示す値である文字パターン尤度が付加される。また文字識別は、第 1 位候補に正しい文字をいつも出力するとは限らないため、下位候補を含む複数個の文字識別候補を出力する。文字列照合では、文字切出しグラフ中から文字パターン尤度の合計が最大となる文字切出しパスを動的計画法により求め、それを文字列認識結果とするのが一般的である[2]。

2.2 文字切出しの課題

個別文字枠がないフィールドに記入する場合、文字同士が接触することがある。たとえば図 2(a) において、'京'の左下点と次の'都'の払いが接触している。また、記入中もしくは記入後に、記入者が記入漏れに気づき、後から文字を追記することがある。たとえば図 2(b) は、'市'を記入し忘れて、後から追記した例である。この場合、'市'と'東'の記入時間が逆転してしまうため、単に記入時間を利用した文字切出しでは正しく切出せない。また戻り書きは既に記入された文字の間に追記するため、接触文字列となりやすい。したがって、高精度化のためには、

課題 1：接触文字列への対応、
課題 2：戻り書きへの対応、

が重要である。

2.3 課題に対するアプローチ

自由文の文字切出し方式として、文字サイズや文字位置、文字間ピッチなど文字パターンの幾何的特徴を利用するアプローチが一般的である。さらに課題 1 の接触文字列に対して時間情報の利用が有効である[3][4]。

本報告では、課題 1 を解決するために、文字サイズや文字位置、文字間ピッチなど文字パターンの幾何特徴と、ストロークやオフストローク(ペン先が紙から離れて動かされた軌跡)の記入時間などの時間特徴、文字識別結果を 1 つの確率モデルで扱い、文字パターンの尤度を計算するアプローチを採用した。

また課題 2 を解決するために、文字パターン作成時に各ストロークの幾何情報と時間情報とを融合利用した。

3. 自由文読取のための文字切出し方式

3.1 プレセグメンテーション

通常文字列は左から右に順々に記入されるため、時間軸に沿って文字パターンを作成していけば、パターンを正しく切出せる。課題 1 の接触文字列も同様に、時間軸基準でパターンを作成すれば、接触箇所のストロークも正しく分離できる。しかし課題 2 の戻り書きでは、追記文字パターンは、時間軸基準では実際より後方に位置付けられてしまう。したがって、戻り書きにも対応するためには、単に時間情報だけでなく記入位置にも配慮した文字パターン作成手法が必要となる。

デジタルペンでは、記入ストロークのサンプリング点情報として、(1)縦方向座標、(2)横方向座標、(3)記入時刻の 3 変量 (x, y, t) が利用できる。これら変量を基準に文字行内のストロークを纏めて、文字パターンを作成すれば良い。提案方式では行内の各文字位置が線状になることに着目し、線形和関数 $f(x, y, t) = w_0 + w_1 x + w_2 y + w_3 t$ によって 3 変量を 1 変量に変換する。この値を指標として行内の全ストロークを整列し、近傍のストロークを順々にマージすることで文字パターンを作成する。

3.2 文字パターン特徴量抽出

対象文字パターンの幾何情報、時間情報、文字識別結果から計 26 個の文字パターン特徴量を抽出する(表 1)。

3.3 文字パターン尤度計算

文字同士が接触していない通常の文字列の切出しには、文字間ピッチなどの幾何特徴が有効である。一方、課題 1 の接触文字列は、通常の文字列で有効な特徴の文字間ピッチが、どの文字パターン間に対しても狭くなり、その差が小さくなる。そのため接触文字列では、文字間ピッチ以外にストローク記入時間やオフストロークが重要となり、そ

† (株)日立製作所 中央研究所, CRL, Hitachi, Ltd.
‡ 北陸先端科学技術大学院大学, JAIST.
(現: 東京農工大学, Tokyo University of Agri. and Tech.)

れら特徴量に重きをおいて文字パターン尤度を計算する方が良い。確率モデルでは、そのような各特徴量の重みを、統計値に基づき算出する。したがって、1 モデルで通常の文字列と接触文字列の両方に対応できる。

本処理では、正解文字パターンの各特徴量の分布が正規分布に従うと仮定し、平均値および分散から確率モデルを構築し文字パターン尤度を算出する。平均値 μ、分散 σ^2 の正規分布において、値 x をとる確率を $P(x|\mu,\sigma^2)$ とする。ある文字パターンにおいて i 番目の特徴量が x_i であった場合、この i 番目の特徴量における文字パターンの尤もらしさ(尤度) L_i は、確率の対数により

$$L_i = \log P(x_i \mid \mu_i, {\sigma_i}^2) = -\frac{1}{2}\log(2\pi{\sigma_i}^2) - \frac{(x_i-\mu_i)^2}{2{\sigma_i}^2}$$

と定義できる。ある文字パターンにおける全特徴量の尤度の総和 $L = \sum L_i$ を、その文字パターンの尤度とする。なお文字識別結果については、その確信度を利用した。

3.4 文字パターンパス探索

本処理は、文字パターン尤度がグラフの各辺に付加された文字切出しグラフを入力とし、グラフ始端から終端を通るパスの中から、文字パターン尤度の合計が最も高くなるパスを動的計画法で決定し、その結果を出力する。動的計画法では、着目頂点に対し、その頂点を終端とするグラフ上の辺を見つけ、その近傍のみで局所的に計算する。それを全頂点に行うことにより、グラフ開始頂点から終了頂点を通る最良のパスを効率良く発見するものである。最終的に、始端から終端までの尤度和が文字切出しパス全体の尤度となり、パスを構成する各文字パターンが切出し対象のパターンとなる。

4. 評価実験

デジタルペンを用い 387 人から表 2の各文字列を収集した。その切出し精度を表 3に示す。なお、文字列単位では文字列中一文字でも切出しを誤ると失敗とした。

実験結果から、接触文字列や戻り書きのサンプルも正しく切出せていることを確認した。一方、句読点やカンマが隣接文字に巻込まれる切出し誤りが存在した。これらは幾何/時間特徴の限界を示しており、文字共起情報などの利用が必要であると考える。

5. おわりに

本報告では、文字パターンの幾何/時間特徴に着目した文字切出し方式を提案した。評価実験から、文字単位の切出し精度が数字で 97.8%、漢字で 91.7%、全字種で 75.6%の結果を得た。

参考文献

[1] http://www.hitachi.co.jp/tegaki

[2] F. Kimura, M. Shridhar, Z. Chen, "Improvements of a lexicon directed algorithm for recognition of unconstrained handwritten words", Proc. of 2nd ICDAR, pp. 18-22, 1993.

[3] T. Fukushima, M. Nakagawa, "On-line Writing-box-free Recognition of Handwritten Japanese Text Considering Character Size Variations," Proc. of ICPR'00, 2000.

[4] 仙田修司, 濱中雅彦, 山田敬嗣, "切り出しパラメータが学習可能なオンライン手書き文字切り出し手法," 信学技法 PRMU97-219, pp.17-24, 1998.

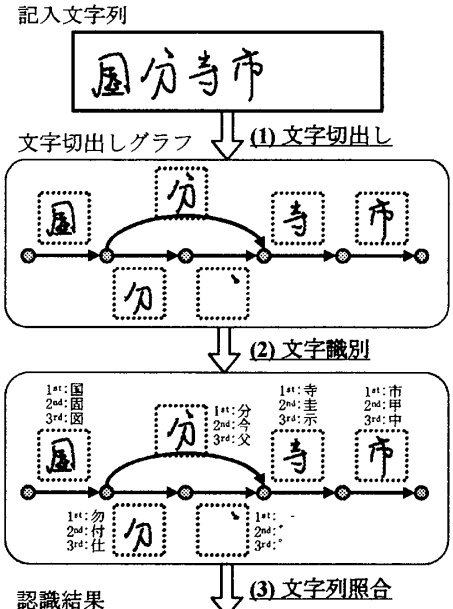

図 1 文字認識処理の流れ

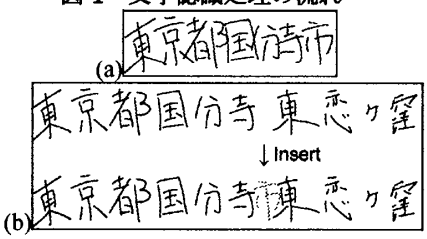

図 2 対応すべきサンプル例

表 1 文字パターン特徴量

特徴量種類	個数
文字パターンサイズ	6
文字パターン位置と重心	4
文字パターン間ギャップ/オフストローク	11
文字パターン内ギャップ/オフストローク	3
各ストローク間距離	1
文字識別結果	1

表 2 評価サンプル

サンプル	文字数	文字列数	行平均文字数	文字列特徴
数字列	20,588	2,077	9.9	数字,ハイフン,カンマからなる
漢字列	1,789	189	9.5	住所表記の地名部分
混合列	50,371	3,580	14.0	自然文や住所表記など,句読点やカンマ等記号含む

表 3 評価実験結果

サンプル	提案方式	
	1 文字単位	文字列単位
数字列	97.8% (20,143)	88.0% (1,827)
漢字列	92.3% (1,652)	71.4% (135)
混合列	75.6% (38,065)	20.6% (736)