書式のない文書画像からの文字パターン列の抽出

2G−5

有田大作†　仙田修司‡　美濃導彦‡　池田克夫‡

†九州大学　‡京都大学

1 はじめに

我々は、スキャナで入力された文書画像を理解し自動的に蓄積する文書画像データベースシステムを構築している。このシステムを構築するためには、文書画像中から文字パターンを抽出し、認識する必要がある。こうした文書画像処理に関する研究は活発に行われている。しかし、これらの研究は、書式のある文書、すなわち、文字パターン列は水平または垂直に並んでいること、段組などによって文書の構造を理解することができること等の制限のある文書のみを対象としている。本稿では、このような制限のない文書から文字パターン列を抽出する手法を提案する。

2 諸定義

以下に本稿における用語の定義を述べる。

文字ストローク 入力画像中の黒画素連結領域。

文字ストローク群 画像上で膨張させることによって結合する文字ストロークの集まり。

文字パターン 1文字に対応する文字ストロークの集合。

文字パターン列 続けて読むべき文字パターンが直線状に並んでいる列。

文字ストローク群の傾き 文字ストローク群を囲む最小の長方形を考え、水平方向から長辺への反時計まわりの角度。

文字パターンの傾き 傾いていない本来の文字パターンからの傾いた文字パターンへの反時計まわりの角度。

文字パターン列の傾き 文字パターン列中の文字パターンの並びの方向への、水平方向からの角度。

3 文字パターン列の抽出

文書は人間が読むことを目的にしているため、書式のない文書画像においても、文字パターンどうしの間隔は、それらが一つの文字パターン列に属するならば狭く、異なる文字パターン列に属するならば広くなっている。本稿で提案する手法はこの性質に着目するとともに、文字パターン認識を利用するものである。本手法は以下の4stepからなる。

step.1 文字ストローク群の構成

画像上で文字ストロークを1ピクセルずつ8隣接で膨張させていく。そして、文字ストロークどうしの結合が起こった時点で、結合の起こった文字ストロークの集まりを文字ストローク群とし、この文字ストローク群についてstep.2以下の処理を行う。このように結合の起こった文字ストローク群のみを随時処理することで、文字パターン間隔の狭い文字パターン列から処理を行うことができる。

step.2 文字ストローク群の傾き決定

文字ストローク群の傾きを、凸閉包を用いた方法で求める。文字ストローク群を一つの図形とみなして作成した凸閉包において、ほぼ平行な2本の長い辺が存在する場合、その辺の傾きを文字ストローク群の傾きとし、step.3の処理を行う。このような辺が存在しない場合、その文字ストローク群は文字パターンどうしの結合でなく、分離文字パターン内での結合であると考え、step.1に戻る。

step.3 文字パターンの切り出し

文字パターンの切り出しには文字パターン認識の信頼度を利用する。これは認識の結果をどの程度信頼して良いかを表す指標であり、次式で定義する。

$$c = \frac{S_2 - S_1}{S_2} \times 100$$

ここで、cは信頼度、S_nは第n位認識候補のスコア(入力パターンと標準パターンの特徴空間における距離)である。また、本研究では方向線素特徴量を用いた文字パターン認識[1]を使用する。

文字パターン列の傾きと文字パターンの傾きの差が90度の倍数になる様な4種類の文字パターンの傾きに対して切り出し処理を行う。これによって、上下隣接、左右隣接の、両方に対応できる。ただし、既に文字パターンの傾きが決定している場合はその傾きについてのみ切り出し処理を行う。

文字ストローク群の傾きに垂直な方向への投影のヒストグラムをとり、このヒストグラムの谷の部分を切り出し位置候補とする。さらに、切り出し候補位置によって切り出される文字ストロークの集合を文字パターン候補と呼ぶ。

Extraction of String of Character Patterns from Unformed Document Images

†ARITA Daisaku,‡SENDA Shuji, ‡MINOH Michihiko,‡IKEDA Katsuo

†Kyushu University,‡Kyoto University

文字ストローク群の中に分離文字パターンの一部のみが含まれている場合があるので、文字ストローク群の端から順に切り出しを行うと文字パターンを正しく切り出せない場合がある。そこで、もっとも面積の大きい文字ストロークを含むような文字パターン候補の中から、文字パターン認識の信頼度の最大のものを、基準文字パターンとする。この基準文字パターンから両端に向かって切り出しを行う。このときも、文字パターン候補の中から文字パターン認識の信頼度が最大のものを選ぶ。

文字パターンの傾きが未決定であれば、各々の傾きについての総合信頼度を計算する。総合信頼度とは、文字パターン列内の文字パターン認識の信頼度の平均である。ある傾きの総合信頼度が他の三つと比べある閾値以上大きければその傾きを文字パターンの傾きとする。その様な傾きがなければ、文字パターンの傾きは未決定のまま処理を進める。ただし、こうして得られた文字パターン列がただ一つの文字パターンから構成されている場合は、文字パターン列の傾きを決定することができないので、step.1 に戻る。

step.4 文字ストローク群の再構成

step.3 の処理で抽出された文字パターン列の傾きの方向に、同じ文字パターン列に含まれるべき文字ストロークが存在する可能性がある。そこで、文字パターン列の方向の近傍の文字ストロークを探す。これによって選ばれた文字ストロークと文字パターン列を統合し、新たに文字ストローク群とする。こうして再構成された文字ストローク群について、再度 step.2、step.3 の処理を行う。この処理を繰り返して、統合すべき文字ストロークがなくなったときの、文字パターン列を出力する。その後、抽出された文字パターン列に含まれる文字ストロークを画像上から削除し、step.1 の処理の続きを行う。

4 実験と考察

図 1 の画像に対して、上述の手法を適用して処理した結果が表 1 である。各誤りについて説明する。

誤り (1) 類似文字によって、1 位候補と 2 位候補の文字パターン認識のスコアの差が小さくなり、文字パターン認識の信頼度が低くなったため。

誤り (2) 文字パターン認識誤りのため。

誤り (3) 文字ストローク群の再構成において、統合すべきでない文字ストロークを統合してしまったため。

文字パターン認識を利用した切り出し処理は、認識結果の信頼度に頼った手法である。類似文字パターンによる信頼度の低下に対して、類似文字パターンを考慮した信頼度を定義する必要がある。

文字ストローク群の再構成では、別の文字パターン列に属するべき文字ストロークを含んだ文字パターン列が抽出されてしまうことがある。このような誤りを防ぐには、抽出された文字パターン列に属する文字ストロークは、文字ストローク群の構成処理 (step 1.) の対象からは外すが、文字ストローク群の再構成処理 (step 4.) の対象からは外さないようにすることが考えられる。このとき、どちらの文字パターン列の一部として抽出するかは、それぞれの文字パターン列内での文字パターン候補の信頼度の大きさによって決定する。

5 おわりに

本稿では、従来の研究が対象にしていなかった、書式のない文書画像からの文字パターン列抽出の手法を提案した。実験によって本手法が有効であることが確かめられた。今後は、4節で述べた問題点に関して研究を行う。

参考文献

[1] 孫寧, 田原透, 阿曽弘具, 木村正行: 方向線素特徴量を用いた高精度文字認識, 信学論 (D-II),Vol.J74-D-II, No.3(1991-3),330-339.

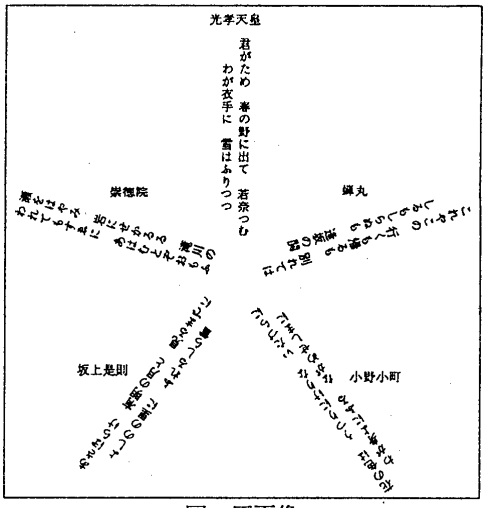

図 1:原画像

表 1:文字パターン認識結果

認識結果 (1 位候補のみ)	結果
君がため春の野に出て若奈つむ	正解
わが衣手に雪はふりつつ	正解
光孝天皇	正解
これやこの行くも帰るも別れては	正解
しるもしらぬも逢坂の関	正解
蝉丸	正解
花の色**^つりにけり**たづらに	誤り (1)、誤り (2)
わが身よにふるとがぬせしまに	誤り (2)
小野小町	正解
あさぼらけ有明のと見るまでに	誤り (3)
よしのの里にふれるしら雪	正解
坂上是則夕	誤り (3)
瀬をはやみ岩にせかるる滝川の	正解
われてもすゑにあはむとぞおもふ	正解
崇徳院	正解