I-019

部分領域を用いた古文書文字列認識に関する研究
Historical Character Strings Recognition Using Partial Area

玉澤 重人† 中山 英久† 和泉 勇治† 加藤 寧† 根元 義章†
Shigeto TAMAZAWA Hidehisa NAKAYAMA Yuji WAIZUMI Nei KATO Yoshiaki NEMOTO

1. はじめに

オフライン手書き文字列認識は，入力画像を切り出して個別文字認識を行う手法が一般的である．文字列画像から個別文字を切り出す処理において，隣接文字間の接触や入り込みにより，個別文字画像にノイズが混入するため，個別文字認識の認識性能が低下する事が知られている．特に古文書文字列では，隣接文字間の接触や入り込みが顕著に見られるため，高精度な切り出しが困難である [1].

本稿では，個別文字画像を抽出する従来の切り出し処理が困難であると考え，古文書文字列認識の新たなアプローチとして，部分領域のマッチングによる文字列認識手法を提案する．古文書文字から疑似的に生成した文字列を用いての認識実験を通じ，提案手法の有効性に関する検討を行った．

2. 部分領域特徴を用いた古文書文字列認識

2.1 辞書の部分領域の作成

辞書は個別文字の学習データについて字種ごとに作成する．文献 [2] では文字画像として 64 × 64pixel の領域を設定し，196 次元の方向線素特徴量を抽出している．一方で本稿は図 1 に示すようにラベル A-1 から A-7 の 64 × 16pixel を領域とした文字画像の部分領域を設定し，方向線素特徴量を抽出している．そしてその部分領域に関する学習データの特徴ベクトルの平均値を辞書とする．つまり部分領域はそれぞれ 28 次元の平均ベクトルを持っており，各字種の辞書は 7 つの部分領域で構成される．

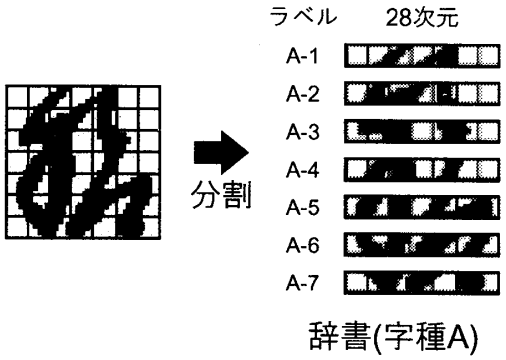

図 1: 辞書

2.2 部分領域毎のラベル割り当て

提案手法が従来手法と最も異なっている点が，本節で提案する部分領域毎へのラベル割り当てアルゴリズムである．入力された未知データについて長方形に正規化し，方向線素特徴量を抽出する．ここで抽出する特徴量は辞書データと同様に部分領域から求まる 28 次元の特徴ベクトルである．そして未知入力の部分領域と，辞書の 7 つの部分領域とのユークリッド距離を算出し，最も距離の小さい上位 N 個の辞書の部分領域のラベルを未知入力の部分領域に割り当てる．このアルゴリズムにより，未知入力の各部分領域に対し，候補となるラベル付けを行う．ラベル割り当ての概念を図 2 に示す．未知入力のある部分領域に対するラベルは，その位置にある字種の一部に似た領域が存在することを示すため，その周辺にはその字種の他の部分領域が存在することが推定される．例えば，A-1 のラベル付けがされた部分領域は，その周辺にラベル A-2 や A-3 が存在している可能性があり，周辺のラベルの分布を調査することで文字認識を実現することが出来る．

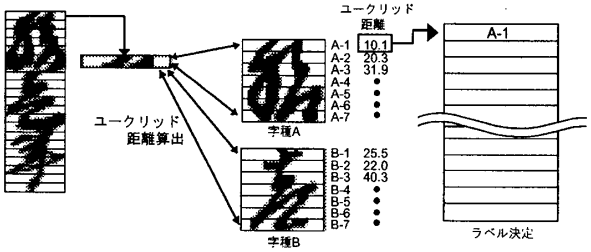

図 2: ラベル割り当て

2.3 ラベルの集中度を用いた文字認識

2.2 節で提案したラベル割り当て手法により，未知データの各部分領域には，辞書データの部分領域に対応するラベルが付加されている．同一字種のラベルが集中している箇所には，その字種の文字が存在している可能性が高いと考えられる．よって本稿では同一字種のラベルの集中度を定量的に評価することにより文字認識を実現する．集中度の算出方法は，各字種の辞書は 7 つの部分領域とラベルにより構成されるため，未知データの 7 つの連続する部分領域を算出単位とする．図 3 に示すように未知データの各部分領域に割り当てられた N 個のラベルに対し，ユークリッド距離の小さい順に単調減少となるような得点を与える．

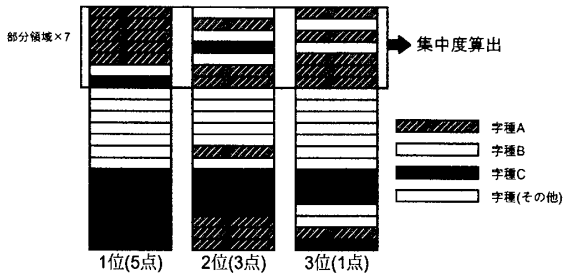

図 3: 集中度の算出

提案手法の部分領域の検出においては，個別文字認識に比べてユークリッド距離を求める際の次元数が小さく，

†東北大学大学院情報科学研究科

誤検出の可能性が高くなる．しかし，一位が誤検出であっても二位，または三位で正読となる辞書の部分領域を検出できれば誤検出を補うことが可能である．

これまで述べた集中度を表したものが図 4 である．集中度の分布を見ると，正読の字種が存在する位置で集中度が高くなっていることが分かる．ここで，字種 A の集中度が連続して高ければ，その位置に字種 A が存在する可能性が高いと考えることができる．そこで集中度の分布において連続して一位を得た字種の得点を集計する．これにより文字列の認識を行うことができる．具体的には分布上部から見て字種 A は連続して 4 回得点が一位となっている 4 つの得点の総和を求め，以降，一位の字種が変わるごとに得点の総和を求めていく．それらを順に並べたものが図 4 の右のグラフである．このグラフに閾値を与え，閾値を越えた字種を認識結果として出力する．本手法では閾値を，集中度の分布において一位を得た得点の平均とした．

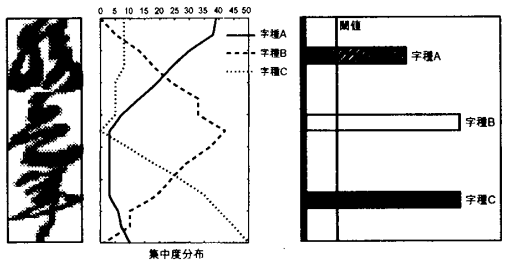

図 4: 文字列認識

3. 認識実験

提案手法による古文書文字列の認識性能を検証するために，古文書文字列認識実験を行う．本実験では古文書文字データベース「HCD1」[1] のうち，表 1 に示す 15 字種 (一字種当たり 200 文字) を使用する．それぞれの字種ごとに辞書を作成し未知データの部分領域の検出を行う．ここで，対象となる 15 字種のみを使用した文字列を実験に十分な数だけ揃えるのは困難なため，個別文字のデータベースから任意に複数の文字を選択して組合せ，疑似的な文字列を作成し実験を行う．作成する文字列の長さは三文字とし，全く同一のデータを持った文字列を得る事がないよう作成した．200 セットの文字列を作成した上で認識実験を行い，文字列の字種と字数を正確に出力したものを正読，それ以外を誤読と判定する．

表 1: 認識対象 15 字種

ツ	一	二	三	四
五	六	七	八	九
十	壱	弐	年	拾

3.1 実験結果および考察

奇数セットと偶数セットの 2 セットに分けて実験を行う．奇数セットとは HCD1 の偶数番の 100 文字によって辞書を作成し，残りの奇数番 100 文字を文字列作成に使用したものである．偶数セットは，その反対に HCD1 の奇数番の 100 文字によって辞書を作成し，残りの偶数番 100 文字を文字列作成に使用したものである．その結果を表 2 に示す．

表 2: 提案手法による古文書文字列の認識結果

	正読数/全データ	正読率
奇数セット	150/200	75.00 %
偶数セット	151/200	75.50 %
トータル	301/400	75.25 %

誤読となった文字列について考察する．まず問題となるのは文字数の誤検出である．本実験の文字列はすべて三文字で構成されているが，二文字，または四文字の文字列と誤検出したものが存在した．この原因として挙げられるのは閾値の設定である．提案手法では集中度の最大値の平均を閾値としたが，閾値を変更することで正読できるであろう文字列がいくつか見られた．閾値の設定は非常に重要であるため，今後その手法について検討する必要がある．

また接触部分以外における部分領域のラベル割り当てにおけるミスもいくつか見られた．この原因としては，ラベルを決定する際に扱う次元数が 28 次元と小さいことが挙げられる．これより当該字種ではない部分領域のラベルを割り当てる可能性が高くなる．解決策としてはユークリッド距離以外の距離尺度について検討し，提案手法に最も適したものを選択する必要がある．

本研究では作成した文字列で実験を行ったが，実際の古文書文字列での認識実験も行う必要がある．その際，文字列における文字の大きさの変動が一番の問題となる．これを解決する手法は非線形正規化を行う事であるが，古文書文字列ではこの文字の大きさの変動が極端な場合があり，正規化の手法は今後大きな課題となる．

4. むすび

本稿では，部分領域特徴を用いた古文書文字列認識手法を提案した．切り出しを用いず，部分領域のマッチングを行うことで古文書文字列を認識できることを示した．実験から提案手法の有効性と誤認識について考察し，今後の課題とするべき点について述べた．トータルで 75.25 %という正読率を得たが，今後更なる改良が必要である．

謝辞

本研究を行うにあたり，古文書翻刻支援システム開発プロジェクトの皆様には古文書文字データベース HCD1 を提供して頂きました．深く感謝致します．

参考文献

[1] 山田奨治, 柴山守, “古文書を対象にした文字認識の研究”, 情報処理, Vol.43, No.9, pp.950-955, (2002)

[2] 孫, 安部, 根元, “改良型方向線素特徴量および部分空間法を用いた高精度な手書き文字認識システム “, 信学論 (D-II), vol.J78-D-II, no.6(1995)