主成分分析を用いた毛筆手書き文字検索

寺沢憲吾　長崎健　川嶋稔夫
はこだて未来大

1　はじめに

古い毛筆文書をデータベース化する需要が増加しているが，毛筆手書き文字は行書や草書といった崩し字体があり，線幅も不安定で安定した細線化を行うことが難しく，従来の文字認識手法を適用することが困難である．

本研究では，文字認識を行うのではなく，文書画像中からある文字列の部分と類似度の高い部分を検出することにより，文字列検索を行う手法を提案する．

2　手法

本手法では，文字画像をまずスリット状に切り出す（図 1 参照）．このように切り出すことにより，文章を画像列としてとらえることができる．

次に，切り出した画像列に対し，Turk ら [1] の固有顔（Eigenface）法にならい，主成分で特徴量を記述する．これにより，画像列は多次元波形データとして表現される．あとは，波形に対してマッチングを行えばよい．

3　実験

3.1　実験条件

実験には図 1 に示す文書画像を用いた．画像サイズは 368 × 1427 ピクセルで，1 文字あたりの解像度はおよそ 60 × 60 ピクセル程度である．文字は 94 文字あり，「蔵人」（16-17, 73-74），「宣百」（27-28, 90-91），「明白也」（50-52, 60-62）の句が 2 回ずつ現れている（数字は文字に出現順につけた番号）．

まず前処理として閾値処理による背景消去と行の切り出しおよび重心算出による中心揃えを行った．次に $\sigma = 3$ のガウシアンフィルタで平滑化した後，画像をスリット状に切り出した．切り出す際のスリットサイズは 9 ピクセル，移動ステップは 5 ピクセルとした．こうして得られた画像列に対し主成分分析を適用し，第 8 主成分までを特徴量に取った．

得られた特徴量に対し，20 ステップ（100 ピクセルに相当）分のウインドウで多次元波形の類似度を算出した．ここでは，次元毎の特徴量の値の差の絶対値の和

$$D(t,t') = \sum_{\tau,d} |I(t+\tau,d) - I(t'+\tau,d)|$$

（$I(t,d)$ は t 番目の画像の d 次元目の特徴量）

をとり，$D(t,t')$ が小さい部分を高類似度とした．

3.2　実験結果

実験結果を図 2 に示す．左の図は文書画像列全体に対して $D(t,t')$ を表示したものである．対応する文字列同士において高い類似度を示している様子が見て取れる．対応する文字列同士以外の部分でも類似度の高い箇所が散見されるが，これは字画が少なく全体に白っぽい文字同士が高い対応を示しているものである．

右の図は，「明白也」の文字列画像に対して，類似度の高い画像上位 3 つを示したものである．正しい文字列が 1 位に検出されている．

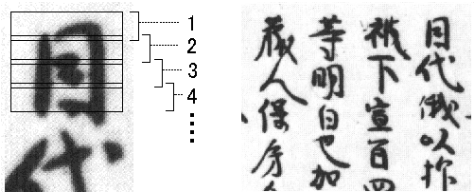

Fig. 1　左：スリット切り出しのイメージ．右：実験に用いた画像の一部

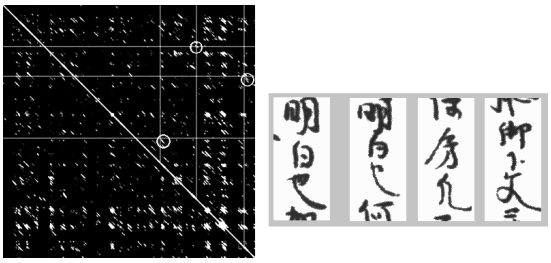

Fig. 2　左：文書画像列全体に対して類似度を行列表示したもの．白い部分が $D(t,t')$ が小さく類似度が高い領域．白丸部が、文字列が対応する箇所．右：（左から）クエリ，検出 1 位，2 位，3 位

4　結論および今後の課題

本手法により文字画像から文字列の検出が可能である．今後は精緻化を目指す．

参考文献

[1] M. Turk, and A. Pentland, “Eigenfaces for Recognition”, Journal of Cognitive Neuroscience, vol.3, no.1, pp.71-86, 1991.