236

中世英文学資料中の印刷文字の認識

山口晃典，内田誠一 (九州大学)
千葉淳一，植竹朋文，松下知紀 (専修大学)

目的：中世英文学資料のデジタルテキスト化

中世英文学資料

西暦1200～1600年代に製作された手稿写本や初期印刷本

デジタルテキスト化(OCR処理)により，

- 差異解析
- 系統解析
- 製作年代
- 製作地域

等の推定が可能に！

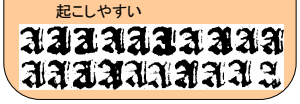
文章の特徴

各資料毎に独特の字体を持つ

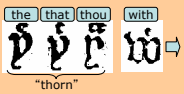
文字の特徴

字形変動
様々な字形変動が存在
インク塗布量や圧力，活字の磨耗や個体差
紙面との2値化の影響

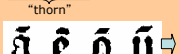
カリグラフ的文字
細かい装飾を施された活字が存在
大文字に多い
構成する線は非常に細い
掠れ，他ストロークとの接合を起こしやすい



特殊文字
余白を調整するための特殊文字



1活字で1単語を表現



後続の“m”や“n”を省略

合字
複数の文字を組み合わせた活字が存在

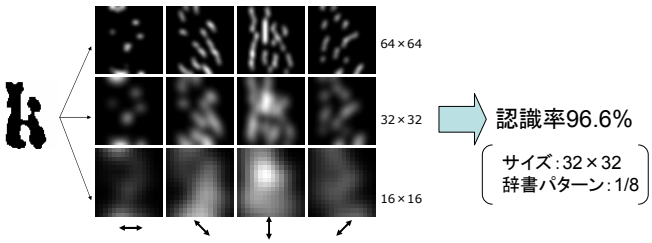
活字間のスペースを自然にするため

現在でも存在するが，異体字と相まって多様に存在

実験

認識実験

特徴：輪郭線の方向成分

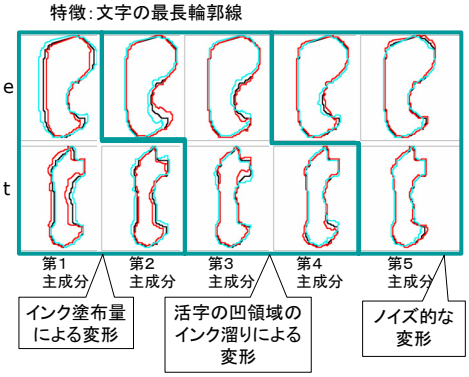


認識率96.6%

サイズ：32×32
辞書パターン：1/8

形状変動の傾向

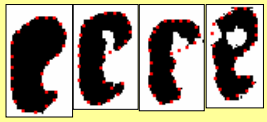
特徴：文字の最長輪郭線



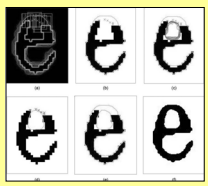
課題とアプローチ

欠損，変形パターンの形状変動

⟶ AAMの利用

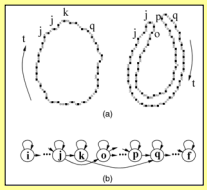


⟶ level set的アプローチ



"Automatic accurate broken character restoration for patrimonial documents "より引用

⟶ HMMの利用



" Integration of Structural and Statistical Information for Unconstrained Handwritten Numeral Recognition "より引用

インタラクティブな操作環境が必要

⟶ GUIの利用