令和元（２０１９）年度　卒業研究報告書概要

課程，学籍番号，氏名	課程：電気・電子情報工学課程　，学籍番号：B183270　，氏名：村井　健
工学分野名：　情報通信システム	指導教員名：　市川　周一

題　　目：和

OpenCV による古文書内の類似文字検索

（英　Detection of the coincidence points of a character image in old documents by OpenCV　）

Abstract

Japanese old documents are hand-written, and often include unreadable or indistinguishable characters. In such cases, it is popular to search the characters in the same or similar shape in the same document, to deduce the original intent of the author. This study investigates the image-processing methods with OpenCV to find the similar characters. Six types of template matching and three types of feature points were examined, where template matching was better than feature point methods. TM_CCOEFF_NORMED was the best out of six types of template matching methods in OpenCV.

概　　要

近年，古文書をデジタルアーカイブ化して公開することが進められている．しかし，古文書は「くずし字」で書かれており，「読めない文字」や「複数文字として解釈可能な文字」が多く含まれる．そのため，古文書中の同じ形の文字を探し，それが同じ文字であると仮定して解読することが多く行われている．そこで，読めない文字とよく似た文字を探す技術が必要になる．

本研究の目的は，古文書画像の中で「似ている文字」を検出する手法の開発である．デジタルアーカイブ化されている画像の処理が必要になるため，OpenCV (Open Source Computer Vision Library) を使用し，テンプレートマッチング，特徴量一致による検出を試みた．古文書画像は「とよはしアーカイブ 橋良文庫 きゝんのこゝろえ」を用いた．

始めに，古文書画像から切り出した「あ」 1 文字をテンプレート画像とし，テンプレートマッチング 6 種類と特徴点一致 3 種類の方法を比較した．最も一致する部分の画像を出力し，白抜きにする．これを繰り返して類似画像の上位 5 つを出力した．その結果，テンプレートマッチングでは複数の検出ができたが，特徴点による方法では切り取った文字しか検出できなかった．また，テンプレートマッチングの 6 種類を比較すると，TM_CCOEFF，TM_CCOEFF_NORMED による方法が「あ」 3 カ所を上位 3 位までに検出していた．TM_CCOEFF に正規化を加えた処理が TM_CCOEFF_NORMED である．

次に，テンプレートマッチングに限定して「あ」以外の文字で検証した．テンプレート画像を全体画像内にあるひらがな 10 種類として，「あ」と同様に上位 5 つの類似画像を検出した．その結果，「は」において TM_CCOEFF，TM_CCOEFF_NORMED のみ 3 カ所と最も多く検出していた．

もう一つの検証として，類似度を閾値として似た文字を検出する方法を試した．この方法では，テンプレートマッチングの検出方法は正規化されている方法に限り，「あ」の検出をして初めて間違った文字の類似度を閾値とした．この類似度をもとに「に」，「食」，「物」の検出を行った．その結果，「に」は TM_SQDIFF_NORMED，TM_CCORR_NORMED では間違った文字数が 65 個と多かったのに対し，TM_CCOEFF_NORMED では間違った文字数が 1 個であった．

これらの検証から，テンプレートマッチングでは似た文字の検出が可能であり，TM_CCOEFF_NORMED による方法が最もよい方法と考えられた．

発表する際の課程を記入　　電気･電子情報工学　　課程　　発表番号　80

（学籍が他課程所属の学生も発表する課程を記入すること）