様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　２８　日現在

機関番号：　１４６０２
研究種目：　基盤研究(B)（一般）
研究期間：　2014～2016
課題番号：　２６２８０１１９

研究課題名（和文）近代書籍自動テキスト化支援環境の構築

研究課題名（英文）Implementation of supporting system and environment for auto-extracting texts from early-modern printed books

研究代表者

城　和貴（Joe, Kazuki）

奈良女子大学・生活環境科学系・教授

研究者番号：９０２８３９２８

交付決定額（研究期間全体）：（直接経費）　　9,200,000円

研究成果の概要（和文）：本研究課題では近代書籍の自動テキスト化を実現するために必要な学習データを効率良く集めるための支援環境を構築した。規格化された現在の書籍用フォントと違い、近代書籍の活版印刷によるフォントにはデータベース等は存在せず、近代書籍から直接画像を切り出し学習データを作成しなければならないが、文字種が1000種類くらいまでは人手でも困難なく収集できるが、2000種を数える頃には困難を極める。そこで不完全ながら学習データを備えた近代文字認識システムを構築し、それに新たな近代書籍を適用し、正しく認識できない未学習の文字を表示させ、その文字種を人間が判断して学習データに追加するシステムを構築した。

研究成果の概要（英文）：In this research, we implemented a supporting system and environment for auto-extracting texts from early-modern printed books. Apart from the current DTP, early-modern printed character recognition requires picture images of early-modern printed books for learning samples. When we collect up to 1000 types characters, the task is not so difficult while when it reaches to about 2000, the task is almost impossible. So we implemented an early-modern printed character recognition system with inefficient learning samples to apply early-modern printed books. The system detects unrecognizable character types to ask user for the correct type. The correctly recognized characters are given to the learning samples so that the recognition system is improved.

研究分野：　パターン認識

キーワード：　近代書籍用OCR　文字認識　特徴量　アンサンブル学習

１．研究開始当初の背景

研究代表者は平成 19 年 12 月より国立国会図書館関西館に非常勤調査員として月 4 回勤務しているが，近代デジタルライブラリのテキスト化の可能性について同館電子図書館課のメンバーと議論を行った．同館所蔵の当時 143,000 冊に及ぶ書籍画像を，青空文庫のように人手でテキスト化を行うのは予算的に不可能であるため，国内の大手 IT 企業に随意契約で近代書籍の自動テキスト化について調査を行わせたところ，既存の OCR 技術では全く役にたたないとの結果であった．この時研究代表者は自身の過去のオフライン手書き文字認識研究からある着想を得た．すなわち近代書籍の多種多様なフォントや旧字体を含めると膨大な種類になる文字セットに対し，これを活字認識とは見なさずに手書き文字と見なすというアイデアである．このアイデアの有効性はすぐに示すことができた．そこでテキスト化されている青空文庫を利用することで，比較的容易に 256 種の漢字を選び出し，手作業での文字切り出しを行った．また学習対象を増やすために文字の自動切り出しを試みたところ近代書籍のルビ除去が非常に困難であることが判明した．出版元が同じであっても時代によって認識できない程全く違うフォントが使われていることも判明した．そして進化的計算を利用したルビ除去手法を開発した．以上のことを踏まえて，近代デジタルライブラリより 36 社 603 冊の書籍を選び，1,000 種類の漢字を異なる出版社･時代から各 5 セット以上作成し，認識実験を行ったところ，学習用データが少なくとも 5 セット確保できる漢字については 98%以上の認識率を得ることを確認した．また当該データセットを作成するために用いたルビ除去では，除去率は 99%以上であった．

２．研究の目的

本課題の目的は，近代書籍の自動テキスト化環境を構築し，国立国会図書館関西館の提供する近代デジタルライブラリの一部を実際に自動テキスト化することである．これまでに申請者らは近代デジタルライブラリの自動テキスト化に関する基礎研究を行なってきたが，これに最低限必要な要素技術は既に確立し，実際に自動テキスト化する際の具体的問題点の把握と問題解決も行っている．そこで近代書籍用活版文字認識システムの学習データをある程度整備し，それ以外の学習データをインタラクティブに生成する支援ツールを開発する．この支援ツールを使うことで近代デジタルライブラリのみならず，新聞雑誌を含む広く近代書籍の自動テキスト化が可能となり，様々な事業を創出し幅広い分野で我国の知的資産価値を上げることが可能となる．

３．研究の方法

本課題ではまず近代書籍用活版文字認識システムの基本学習データセットを作成し，同時に拡張学習データ収集支援ツールの開発を行う．基本学習データセットの作成には青空文庫を利用し，どの漢字がどのタイトルのどの辺に出現するかを使って，近代デジタルライブラリの該当箇所漢字画像を手動で切り出す．拡張学習データ収集支援ツールは Web アプリケーションの形を取り、近代書籍用文字認識エンジンと学習データ DB とで構成され､ユーザは Web ブラウザから作業を行う。作業の流れは次のとおりである。まず新規の画像データを支援システムに入力する。支援システムは入力された近代書籍画像から文字切り出しを自動的に行った後、複数の特徴抽出を行い、複数の識別機で認識を行う。認識結果のスコアが低い場合、誤認識である確率が高いので、ユーザにどの文字が誤認識の可能性が高いかを画像とテキストで提示する。ユーザはその結果を目で見て正しく認識していない場合にはその文字の種類を入力する。支援システムは誤認識した文字を新たに学習データとして学習データ DB を増強する。このように基本学習データセットと支援ツールを使って学習データ数を増やし，適宜新規学習を行わせ支援ツールの認識率を徐々に上げていく．この作業は研究期間終了まで続け，対応文字種を少しでも多くする．

４．研究成果

雑誌論文①では３．で述べた基本学習データと近代書籍用文字認識エンジンでの認識結果を報告した。

雑誌論文②では近代書籍用文字認識エンジンにおける特徴抽出手法を三種類に増やし、それがどのように識別結果に影響を与えるかを分析した。その結果、アンサンブル学習

を使うことで近代書籍用文字認識エンジンを実際の利用に耐えうる精度に改良できる可能性について言及した。
雑誌論文③では基本学習データと近代書籍用文字認識エンジンを使ってWebブラウザから学習データを効率良く収集することができる学習支援システムを構築したことについての発表を行った。
雑誌論文④ではある程度の規模の学習セットに対して我々の提案している近代書籍用文字認識手法が有効であることを示した。
雑誌論文⑤では文字切り出しを行う際のルビ除去は現在広く使われている手法では近代書籍に適用できないことを示し、進化計算を使った新たな手法を提案した。

本研究課題では、学習データ収集支援システムを利用することで必要な学習データが揃うという前提で計画を立てていた。しかしながら実際の収取を行っていくと、どの書籍にも見つからない文字が非常に多くあることが判明した。すなわちZIPの法則である。この問題に対処するために、学会発表②と③では進化計算を用いて既知の文字からなる学習データを学習させることで、同じフォントセットの未知の文字を自動生成する研究に着手したことを報告した。このサブテーマはあまり良い結果を出さなかったが、最終年度に進化計算ではなくディープラーニングを用いたコンボリューショナルニューラルネットで目覚ましい成果を上げることができたので学会発表①を行った。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）
①粟津 妙華, 上坂 和美, 高田 雅美, 城和貴：近代書籍を対象とした多フォント漢字認識, 情報処理学会論文誌数理モデル化と応用, 査読あり, Vol.9(2), pp33-40 (2016).
②Kazumi Kosaka, Kaori Fujimoto, Yu Ishikawa, Masami Takata, Kazuki Joe: Comparison of Feature Extraction Methods for Early-Modern Japanese Printed Character Recognition, 2016 International Conference on Parallel and Distributed Processing Techniques and Applications, Final Edition, 査読あり, pp.408-414 (2016).
③Kazumi Kosaka, Taeka Awazu, Yu Ishikawa, Masami Takata, Kazuki Joe: An Effective and Interactive Training Data Collection Method for Early-Modern Japanese Printed Character Recognition, The 2015 International Conference on Parallel and Distributed Processing Techniques and Applications, Vol.1, 査読あり,pp. 276-282 (2015).
④Taeka Awazu, Manami Fukuo, Masami Takata, Kazuki Joe: A Multi-Fonts Kanji Character Recognition Method for Early-Modern Japanese Printed Books with Ruby Characters, 3rd International Conference on Pattern Recognition Applications and Methods, 査読あり, pp. 637-645 (2014).
⑤粟津 妙華, 高田 雅美, 城 和貴：活字データの分類を用いた進化計算による近代書籍からのルビ除去, 情報処理学会論文誌数理モデル化と応用, 査読あり, Vol.8(1), pp72-79 (2014).
〔学会発表〕（計 2 件）
①竹本 有紀, 上坂 和美, 石川 由羽, 高田雅美, 城 和貴：近代書籍用フォントの自動生成, 情報処理学会数理モデル化と問題解決研究会, 2017-MPS-112(15),1-6 (2017-2-27).
②大坂智葉, 粟津妙華, 石川由羽, 高田雅美, 城和貴: GPを用いた活字風手書き文字の生成, 進化計算学会進化計算シンポジウム 2015 (2015-12-20).
③岩田彩, 上坂和美, 粟津妙華, 石川由羽, 高田雅美, 城和貴: 近代書籍用 OCR のための学習用特定フォントセットの自動生成手法, 情報処理学会数理モデル化と問題解決研究会, 2015-MPS-105(10), 1-6 (2015-09-22).

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
城和貴　　（JOE, Kazuki）
奈良女子大学・生活環境科学系・教授
研究者番号：90283928
(2)研究分担者
高田雅美　（TAKATA, Masami　　）
奈良女子大学・生活環境科学系・講師
研究者番号： 20397574

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
木目沢司（KIMEZAWA,Tsukasa）
国立国会図書館関西館・電子図書館課・書士