様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月　 ６ 日現在

機関番号：１２６０５
研究種目：基盤研究(B)（一般）
研究期間：2012～2015
課題番号：２４３０００９５

研究課題名（和文）アジア文化圏の古文書アーカイビングのための基盤構築

研究課題名（英文）Preparing a test bed for digital archives of historical documents in Asia

研究代表者
中川　正樹（NAKAGAWA, Masaki）

東京農工大学・工学（系）研究科（研究院）・教授

研究者番号：１０１２６２９５

交付決定額（研究期間全体）：（直接経費）　　12,900,000円

研究成果の概要（和文）：本研究では，アジア文化圏の古文書を電子的にアーカイブするための基盤技術を研究した．ベトナムのチュノム文書を解読するためには，4万字種にも及ぶチュノム文字認識システムを作成した．また，奈良平城京などから出土する木簡を対象に，ノイズ除去や切出しを含むアノテーションツールを開発した．このアノテーションツールは，奈良文化財研究所で，破片になった木簡を整理するために利用されている．　言語の適用範囲を拡大する研究を継続している．

研究成果の概要（英文）：This research studied a test bed for digital archives of historical documents in Asia. For Vietnamese Chunom documents, we made OCR of more than 40,000 character categories. For wooden tablets excavated from the Heijo palace site, we made an annotation tool including noise reduction and extraction. This annotation tool is availed for archiving an enormous amount of broken pieces of wooden tablets. We are extending the research to other languages in Asia.

研究分野：情報工学

キーワード：古文書　アーカイブ　画面処理　言語処理　文字認識

１．研究開始当初の背景

(1) 古文書電子化の重要性，アジア圏の字種の特殊性

歴史文書は，世界の文化や歴史を研究する上で重要な資料である．それらの中には，劣化や破損・汚損により解読が容易でないものや，現在使われない文字によるものも多く含まれ，情報技術を活用した解析や電子化が望まれている．

特にアジアには，世界四大文明のうちの三つが含まれ，多くの重要な歴史文書が存在する．それらは，字種に大きな特徴を持つ．日本語，中国語，韓国語は，部首を組み合わせて数千以上の字種を構成し，分かち書きをせず，縦書き横書きが混在するという特徴を共有している．インドに発する東南アジア言語においても，子音字に母音字を組み合わせるため，実質的な字種は多い．中近東では，独特の字種からなるアラビア文字が古くから使われている．このようなアジアの字種の特殊性は，手書き文書の認識や解析の際に，文字の切出し，認識，言語処理において共通する問題を生じ，欧米言語に有効な手法（隠れマルコフモデルなど）だけでは解決できない．

(2) 欧米の組織的取り組み，アジアの遅れ

歴史文書処理の研究は世界的に活発化しており，図書館学，歴史学，パターン認識などの分野単独／連携で国際会議が多数開催されている．特に欧米では以下の組織的取り組みが進んでいる．

(a) 米国における古文書の電子図書館化：L. D. Paulson：Finding ways to read and search handwritten documents, IEEE Computer, 38, 3 pp.22-24 (Mar 2005)．など

(b) EU による歴史文書の電子化基盤構築：H. Balk and L. Ploeger: IMPACT: working together to address the challenges involving mass digitization of historical printed text, OCLC Systems & Services,25,4,pp.233-248(2009)など

一方，アジアにおいては，個々の国立図書館などで電子化の取組みが行われ始めているものの，電子画像に変換しているだけのレベルで，欧米に比べて大きく遅れをとっている．

２．研究の目的

アジア文化圏の様々な古文書の電子アーカイブのために，統一的処理技術を確立するするとともに，システムの基盤となるライブラリ群を構築し公開する．(1) 木簡や竹簡などの媒体に特有な劣化や汚損・破損に対応できる画像処理，(2)言語に共通する文書解析，(3) 言語依存の文書・文字認識，(4) 上記機能によるタグ付けと手書きアノテーションの付与と活用，の 4 つの基盤技術を確立し，ライブラリとして公開する．言語からの独立性を高め，言語依存の部分についても，言語ごとに開発するのではなく，言語共通のメタな枠組みを提供する．また，複数の考古学者がこれらの機能を利用して，遠隔協調で解析し議論できる環境も整備する．本研究により，文化的・歴史学的にきわめて重要な，アジア圏全体の歴史文書の電子化推進に貢献する．

３．研究の方法

① 媒体に特有な画像処理方式の確立とライブラリ化
② 言語に共通な文書解析手法の確立とライブラリ化
③ 言語依存の文字認識と文書解析の確立とライブラリ化
④ 上記機能によるタグ付けと手書きアノテーションの付与と活用

これらの達成のために作業班を分けるが，十分な連携をとって研究開発を進める．2 年目までに一応のシステムを作成し，あとはスパイラル法により，各手法の完成度を高め，ライブラリを更新していく．主軸にベトナムのチュノムによる古文書を取り上げることにより，日本語あるいは英語文書に慣例として

使われてきた手法を見直し，言語独立の機能を高めたいと考えている．

４．研究成果

24 年度は，(1)，(2) (3) で基礎的研究を行い，特に(2)において，縦横へのヒストグラムの解析やボロノイ図，ハフ変換などによる行や文字への分割，ラベリングやモルフォロジによるノイズ除去，傾き補正や正規化などを開発し，また(3)において，古文書から切り出した文字パターンをクラスタリングしてラベル付けを行い，文字パターンデータベースを構築した．さらに，この段階のクラスタリングは精度が低いので，対話的に修正ができるシステムにした．そして，文字パターンデータベースの大半を学習パターンとして，文字認識エンジンを開発した．

25 年度では，それらの改良に加えて， ベトナムのチュノム文書を解読するために，4 万字種にも及ぶチュノム文字認識システムのプロトタイプを作成した．また，奈良平城京などから出土する木簡を対象に，アノテーションツールを開発した．このアノテーションツールは，奈良文化財研究所で，破片になった木簡を整理するために利用されている．また，破片を自動的に組み合わせる研究も行っている．

アノテーションには，文字と図を記入することが多いので，それらを分離する方式を開発し，日本語と英語の大規模データベースでその効果を実証した．これは言語によらず利用可能である．さらに，ベトナムのチュノム文書に対して，英語あるいはベトナム語でアノテーションを記入し，認識させてコード化できるオンライン手書き英語認識システムと手書きベトナム語認識シスシムも作成した．これらの成果は，国際学術誌や国際会議議事録に採録された．ベトナムの大学との国際共同研究が本格化した．

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３件）

① Truyen Van Phan and Masaki Nakagawa, Combination of Global and Local Contexts for Text/Non-text Classification in Heterogeneous Online Handwritten Documents , Pattern Recognition, 査読有, Vol. 51, pp. 112-124 (2016. 3), DOI URL:10. 1016/j. patcog. 2015. 07. 012.

② Truyen Van Phan, Kha Cong Nguyen, and Masaki Nakagawa, A Nom historical document recognition system for digital archiving, International Journal on Document Analysis and Recognition (IJDAR) , 査読有, Vol. 18, pp. 1-16 (2015. 12), DOI URL:10. 1007/s 10032-015-0257-8.

③ Phan Van Truyen, 中川正樹, 馬場基, 渡邊晃宏, 木簡画像集録システムの設計と実現, 日本情報考古学会, 査読有 , Vol. 19, No. 1 ・ 2, pp. 1-12, 2013 (2013. 5).

〔学会発表〕（計１４件）

① Kha Cong Nguyen, Nakagawa Masaki: Text-Line and Character Segmentation for Offline Recognition of Handwritten Japanese Text, IEICE technical report, Vol. 115, No. 517, pp. 53-58, 産業技術総合研究所, 臨海副都心センター, 東京都江東区 (2016. 3. 24).

② Hung Tuan Nguyen, Cuong Tuan Nguyen, Pham The Bao , Masaki Nakagawa: Preparation of an Unconstrained Vietnamese Online Handwriting Database and Recognition Experiments by BLSTM, IEICE technical report, Vol. 115, No. 517, pp. 59-64, 産業技術総合研究所, 臨海副都心センター, 東京都江東区(2016. 3. 24).

③ 稲谷壮一郎, 中川正樹：オンライン手書き文書に対する BLSTM ニューラルネットワークによる文字・図形分離の再検討, 信学技報, Vol. 115, No. 517, pp. 65-70 産業技術総合研究所, 臨海副都心センター, 東京都江東区(2016. 3. 24).

④ Cuong Tuan Nguyen and Masaki Nakagawa: An Improved Segmentation of Online English Handwritten Text using Recurrent Neural Networks, in Proceeding of the 3rd IAPR Asian

Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia (2015.11.4).

⑤ Kha Cong NGUYEN, Truyen Van PHAN, Masaki NAKAGAWA: A System to Annotate and Cluster Pieces of Mokkan, in Proceeding of the 2015 Fourth ICT International Student Project Conference, 東京農工大学,東京都小金井市 (2015.5.24).

⑥ Hung Tuan NGUYEN, Cuong Tuan NGUYEN, Pham The BAO, Masaki NAKAGAWA: A Vietnamese Online Handwriting Database, in Proceeding of the 2015 Fourth ICT International Student Project Conference, 東京農工大学,東京都小金井市 (2015.5.24).

⑦ Cuong Tuan Nguyen and Masaki Nakagawa: An Improved Segmentation of Online English Handwritten Text using Recurrent Neural Networks , in Proceeding of the 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia (2015.11.4).

⑧ 中川正樹,Phan Van Truyen:失われた言語チュノムによる古文書の完全電子化に向けて,日本情報考古学会講演論文集（第33回大会）,Vol. 13, pp.84-88,東京農工大学,東京都小金井市(2014.9.27).

⑨ Cuong Tuan Nguyen, Bilan Zhu and Masaki Nakagawa: A semi-incremental recognition method for online handwritten English text, Proc. 14th International Conference on Frontiers in Handwriting Recognition (ICFHR2014), Crete, Greece, pp.234-239 (2014.9.2).

⑩ Truyen Van Phan, and Masaki Nakagawa: Text/Non-Text Classification in Online Handwritten Documents with Recurrent Neural Networks , Proc. 14th International Conference on Frontiers in Handwriting Recognition (ICFHR2014), Crete, Greece, pp.23-28 (2014.9.2).

⑪ Truyen Van Phan, and Masaki Nakagawa: Construction of a Text Digitization System for Nôm Historical Documents, Proc. International Conference on Digital Access to Textual Cultural Heritage (DATeCH2014), Madrid, Spain, pp.65-70 (2014.5.19).

⑫ Cuong Tuan Nguyen, Bilan Zhu and Masaki Nakagawa:semi-incremental recognition method for on-line handwritten Japanese text, Proc. 12th International Conference on Document Analysis and Recognition (ICDAR2013), Washington D.C.,USA,(2013.8.26).

⑬ Truyen Van Phan, , Hajime Baba, Akihiro Watanabe, and Masaki Nakagawa:A Re-Assembling Scheme of Fragmented Mokkan Images, Proc. 2nd International Workshop on Historical Document Imaging and Processing (HIP'13) , Washington D.C.,USA,(2013.8.23).

⑭ Truyen Van PHAN, Hajime BABA, Akihiro WATANABE and Masaki NAKAGAWA:MokkAnnotator - A System for Archiving Mokkan Images, Proc. 16th International Graphonomics Society Conference (IGS2013) 東大寺,総合文化センター,奈良県奈良市,(2013.6.11).

〔図書〕（計　０　件）

〔産業財産権〕
○出願状況（計　０　件）
○取得状況（計　０　件）

〔その他〕
ホームページ等
なし

６．研究組織
(1)研究代表者
中川　正樹（NAKAGAWA, Masaki）
東京農工大学・大学院工学研究院・教授
研究者番号：10126295

(2)研究分担者
朱　碧蘭　（ZHU, Belin）
東京農工大学・大学院工学研究院・助教
研究者番号：50466918

斉藤　隆文（SAITO, Takafumi）
東京農工大学・大学院工学研究院・教授
研究者番号：60293007

堀田　政二　（HOTTA, seiji）
東京農工大学・大学院工学研究院・准教授
研究者番号：90346932

(3)連携研究者
なし