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

令和　２ 年　 ７ 月 １３ 日現在

機関番号：３３９２１
研究種目：若手研究(B)
研究期間：2017～2019
課題番号：１７Ｋ１２７９０

研究課題名（和文）クラウドソーシングを用いた効率の良いデジタル翻刻手法の開発

研究課題名（英文）a Study of Efficient Human-assisted OCR of Japanese Books

研究代表者
池田　光雪（IKEDA, Kosetsu）

愛知淑徳大学・人間情報学部・講師

研究者番号：１０７７９６０６

交付決定額（研究期間全体）：（直接経費）　　1,100,000円

研究成果の概要（和文）：アーカイブサミット2017や三田図書館情報学会などの講演における，図書館やアーカイブ関係者との議論を通じ本研究で必要とされる複合的なタスクデザインがどうあるべきかについて検討を行い，そのデザインを完成させた．また諸般の事情により論文として成果の公表までには至らなかったが，マイクロタスク型クラウドソーシングを用いたシステムの設計を行うことができた．

研究成果の学術的意義や社会的意義
デジタルアーカイブが強く推進されており，様々な文書が電子化されていくなかで，従来の人手やOCR単独の処理に依らない，マイクロタスク型クラウドソーシングによるデジタル翻刻の新たな方向性を示すことができた．今後，サービスとして定着するところまでこの成果を発展させることができれば今後の情報基盤として有用なものになると考えられる．

研究成果の概要（英文）：Through discussions with the participants in the lectures, the task design needed in this study was completed. Unfortunately, the results were not published as a paper due to various reasons, we were able to design a system using microtask-type crowdsourcing.

研究分野：情報工学

キーワード：クラウドソーシング　デジタル翻刻　マイクロタスク

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

本研究の背景として，画像化された資料の利活用がある．オープンデータやオープンサイエンスの潮流による後押しも受け，これまで各機関が管理してきた貴重書や，あるいは著作権保護期間が満了した資料を対象とした画像化及びその公開，いわゆるデジタルアーカイブは広く行われるようになった．ここで，通常デジタルアーカイブにおいてはページを撮影しそれを高精細な画像として公開しているが，それに加えて記載された内容をデジタル翻刻（文字起こし）して文字データとして同時に提供することができれば，単語の使用回数のような定量的な解析や全文検索，機械による読み上げが可能になるといった恩恵を得ることができる．

しかし，そのデジタル翻刻を行うためには人手あるいは OCR による作業が必要となる．人手による作業は間違いを起こしにくく高品質な結果を出すことのできる一方で非常に膨大な手間と時間がかかり，OCR はわずかな時間で終えることができる反面，使用文字数が欧米に比べて膨大となる日本語に対しては十分な質の結果を得ることができず，結局人手による綿密な修正が必要になるという問題がある．

これらを背景として，人手と OCR，両者の強みを活かしたデジタル翻刻プロジェクトが特に欧米を中心として多数発足している．近年では日本においても古典籍を中心として複数のプロジェクトが立ち上げられている．これらは OCR 結果のデータと元の画像ファイルを同時に確認できるように Web 上で公開し，OCR の結果に間違いがあればそれを有志が修正するという手法が基本的である．この手法では，原文の解読にある程度の能力が必要となるくずし字で書かれた資料に対しても短時間で大量のデジタル翻刻ができているなど，従来の専門家のみに比べたデジタル翻刻に比べれば多大な成果を挙げていると言える．

一方で，この Web 上で不特定多数の人の力を借りてデジタル翻刻を推し進める仕組みは，ごく一部の熱心な作業者が作業の大多数を行うという傾向が強く見られ，プロジェクトの完遂にはそのような作業者を獲得できるかという点に課題が見受けられる．

２．研究の目的

本研究計画では不特定多数の人々に作業を委託するクラウドソーシングに着目し，作業者間の負荷を分散させ，かつ効率良くデジタル翻刻を行う手法の構築を行うことを目的とする．

特に，行うべき作業を細かい単位に分解させるマイクロタスク型クラウドソーシングを採用することで，新たなデジタル翻刻手法の提案を目指す．これにより，従来手法と比べ１回の作業単位が非常に短期間となり，アプリケーションを通じてパソコン以外の媒体でも様々な形態で実行が可能になることで，参加へのハードルが下がることが期待される．さらに，タスクを多段階に組み合わせる設計を行うことにより，単一のタスク設計よりもロバストなタスク設計を目指す．

３．研究の方法

マイクロタスクを用いてデジタル翻刻を行うために，まずそのタスクデザインを検討する．タスクデザインはタスク結果の質と必要なタスク数を決める重要なものであるが，複数のタスクデザインを組み合わせることにより質を保ちつつもタスク数を減らすデザインを目指す．これにより，OCR が文字と誤判定したが実際にはイラストなどで文字起こしを行う必要がない箇所を翻刻させようとする無駄を省くことができる．

さらに，単に OCR 結果を１文字１文字修正させるタスクを実行させるのではなく，タスク結果から推測される OCR 精度から１度に修正対象とする文字数を動的に変化させることで，よりすくないタスク数で翻刻の完了が行えるような設計を行う．

また，マイクロタスク型クラウドソーシングでは Web を使ったシステムのようにパソコン上でしか作業ができないわけではなく，スマートフォンを始めとした様々な媒体でタスクを行うことが可能である．これにより，ロック画面の解除時にタスクを行わせるようにするアプリといった先行研究も取り込むことが可能であるが，それぞれの媒体に応じたタスクの最適化も検討する．

これらで考案したタスクデザインに基づきシステムを実装し，実際にデジタル翻刻を行うことでデータの収集とタスクデザインの評価を行う．

４．研究成果

アーカイブサミット 2017 や三田図書館情報学会などでの講演を通じ，アーカイブ関係者及び図書館関係者らと本研究課題に関する議論を行い，本領域の動向への理解を深めると共に，タスクデザインに必要となるニーズを再確認することができた．これにより，理論的な計算量も含めた多段階のタスクデザインは完成した．

しかし，諸般の事情によりそのタスクデザインを用いたシステムの実装と，それにより得られ

たデータを分析し，論文として成果の公表を行うことまでは至らなかった．今後の展望として，システムの実装を完遂して公表することが挙げられる．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　2件／うち国際学会　0件）

１．発表者名 橋本雄太，亀田尭宙，池田光雪
２．発表標題 デジタルアーカイブの情報技術
３．学会等名 アーカイブサミット2017（招待講演）
４．発表年 2017年

１．発表者名 池田光雪
２．発表標題 マイクロタスク型クラウドソーシングを活用した図書館における取り組み
３．学会等名 三田図書館情報学会（招待講演）
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考