科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月　 ４ 日現在

機関番号：１３４０１
研究種目：基盤研究(C)
研究期間：2011～2013
課題番号：２３５００１１６

研究課題名（和文）文書の置き方に依存しない文書情報入力システムの構築

研究課題名（英文）Development of a document processing system without depending on document orientation

研究代表者
長谷　博行（Hase, Hiroyuki）

福井大学・工学（系）研究科（研究院）・教授

研究者番号：９０１４２２７３

交付決定額（研究期間全体）：（直接経費）　　2,400,000円、（間接経費）　　720,000円

研究成果の概要（和文）：本研究は、膨大な学術雑誌や書籍の内容を電子化することを目的として、高解像度カメラで取得した文書画像の情報処理システムの構築を目指した。特に、文書の置き方に制限を設けない手軽な入力作業を可能にする。本研究では、具体的には、電子化されてない大量の古い年代の論文誌を半自動入力するためにページ捲り器を用いて、冊子を裁断することなく見開きページ画像をカメラから入力し、レイアウト解析、角度推定が可能な文字認識を適用することによりデータベースを構築する簡易なシステムを構成した。また、1文字毎の向きを推定することが可能で複数フォントの日本語約3000カテゴリに対応可能な文字認識方式を開発した。

研究成果の概要（英文）：This project aims at developing a document information processing system which copes with arbitrarily inclined documents, using a high resolution camera. This system is used to the arrangement and the retrieval of a large amount of document information database. A new system using the page turning machine is developed in order to input a lot of academic journals, this system constructs a small size of searchable paper database. This project is divided into two problems, that is, the one is to develop a document processing algorithm. The other is to develop the character recognition method which can estimate the orientation angle of the character and can cope with about 3000 Japanese Kanji categories with Mincho font and Gothic font.

研究分野：総合領域

科研費の分科・細目：情報学・メディア情報学・データベース

キーワード：文書画像処理　レイアウト解析　回転文字認識　文献データベース　ページ捲り器

１．研究開始当初の背景

CiNii,J-STAGE などの論文データベースはいくつか存在する。JST 国内収集誌のうちアグリゲータサービスに収録されている電子資料は，延べ 9,639 誌であった。しかし、全文検索できる論文の他に、タイトル、タグ、紀要等の条件検索のみの論文も多い。JST 国内収集誌の電子化状況調査報告(2012)によると、2012 年 2 月時点での JST 国内収集誌の内、全文検索（抄録や要約のみでなく本文が電子化されたもの）可能な学術誌・学会誌は 2,478 誌のうち 1,460 誌である。研究報告・技術報告では 3,138 誌のうち 1,836 誌である。つまり、電子化された学術誌・学会誌は 59 %と非常に少ないため、目的の論文を検索する際に約 59 %の論文を見逃してしまうことになる。これは研究活動において、マイナスである。

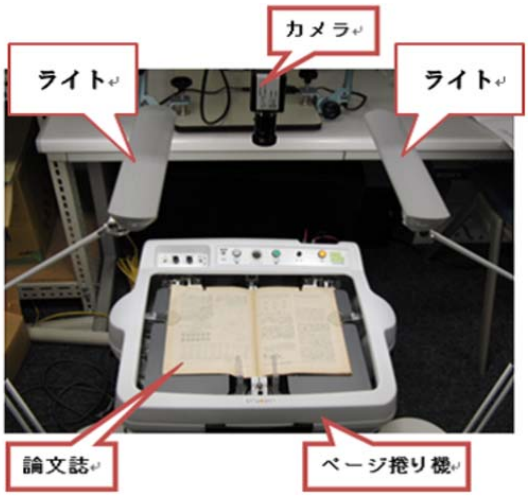
このような背景から、本研究ではデータベース化されていない文献に焦点を当てる。大学や研究所の研究室や図書室ではそのような文献が多く存在する。これらの論文をパーソナルなシステムにより簡易に電子化、データベース化できるようなシステム作成を目標とする。

２．研究の目的

最近のインターネット検索の便利さから、データベースに漏れている文献は参照されない状況が生じている。データベースに登録されている論文は PDF 形式で保管されているが、少し古い論文は画像として登録されているものが多く、検索項目は限られる。さらに古い論文誌は登録すらされていない。そこで、本研究では、これら大学の研究室に保管され、埃がかぶっている論文誌群を電子化し、小規模なデータベースとして利用することを目指す。ただし、紙面を裁断せずに紙面画像を入力する。そのために、障害者や高齢者が本を読むために使う安価なページ捲り器を利用する。この装置により 1 冊の論文誌を自動でページを捲り、高解像度カメラにより全ての見開きページを撮影することができる。この見開きページを各ページに分離し、レイアウト解析し文字認識を適用した後、論文固有の知識からキーワードである［論文］、［ショートノート］、[サーベイ論文]、[研究速報]、[招待論文]を検出し、ページ画像列をそれらの個別論文毎に分離し、可能な限り多くの項目を抽出してデータベースに登録するシステムを目指す。

しかし、ページ捲り器の冊子の置き方の制限や動きに起因するページの傾きや、研究会技報に見られる横向き文字のレイアウトにも対応できるようにするには、文字の回転角度が推定可能な認識方式の開発が必要である。故に、本研究を二つの問題に分け、ある程度満足できる性能を得た時点で統合する。

(1)一つは論文誌画像のカメラ入力からデーベース構築までの処理法の検討と実現で



ある。認識対象は古い冊子が多いため、二値化、ノイズ除去、レイアウト解析などにいくつかの検討課題がある。

(2)他のひとつはページの歪みや文書の置き方に依存しないようにするためには文字認識と同時に文字の回転角を推定する必要がある。これまでもこの研究を行ってきたが認識対象は英数字の 62 文字種のみであり、これを日本語文字約 3000 文字種に拡張する必要がある。さらにいくつかのフォントにも対応する必要がある。これら二つの課題について研究開発を行った。

３．研究の方法

(1) ページ捲り器と高解像度カメラを用いた文書処理方式

文書画像処理のためのアルゴリズムはこれまでに多数提案されているが、文書の状態により独自の工夫が必要である。本研究では、論文誌のページを自動的に捲る装置であるページ捲り機に論文誌を設置する(上図)。論文誌は上方に固定した照明により一定の明るさを確保する。高解像度カメラ及びページ捲り器の制御用インタフェースを作成して、ページ捲り器を PC から制御する。ページを捲る毎にページ捲り機の上方に固定した高解像度カメラで撮影（4384×3288）し、ページ分割、二値化、レイアウト解析、角度推定可能な文字認識を適用することにより、全文検索可能な論文データベースの構築を目指す。この手法では論文誌を裁断することなく、電子化できる利点がある。まず、見開き 2 ページ画像をページ境界の影を利用してそれぞれのページ画像に分割する。次に、各ページ画像に対し二値化を行う。その際、ページ面がフラットではないため明るさにムラが生じ、固定閾値では潰れや掠れが生じる。そこで最適な局所二値化法を検討する。また、二値化の際に発生するごま塩ノイズは膨張・収縮処理で取り除くことができるが、i の上の点も消去される。このような副作用が起きないよう工夫が必要である。ページ画像の傾きは本来なら課題(2)の回転文字認識の成果を待つところであるが、横書きを仮定した代替手法として、黒画素を横方向に集積し

たプロファイルの形状から補正した。レイアウト解析は、ページ画像の黒画素を縦横方向に集積したプロファイルから再帰的領域分割を行う。また、文字認識には、文字の向きが正しいと仮定し、代替ソフトを Google のフリーOCR エンジンの Tesseract を用いる。これを抽出した文字行に適用する。ここまでの処理で、文字行を取り出して文字認識結果が得られる。

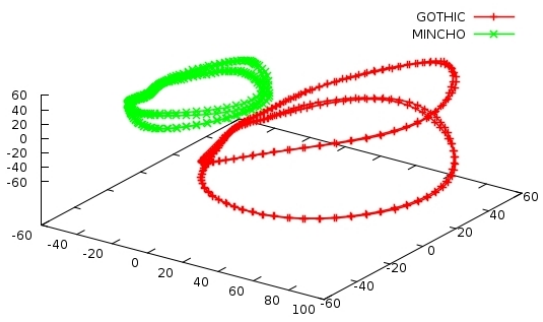
次に、ページ画像列から論文単位に区切るには、論文の先頭ページの識別が必要である。検出方法としては、レイアウト構造の違いに基づく方法もあるが、本課題では、対象論文の先頭ページに出現する論文固有シンボルを検出する。電子情報通信学会論文誌の場合、先頭ページ検出率を［論文］、［ショートノート］、[サーベイ論文]、[研究速報]、[招待論文]に対して実験的に検討する。分割された各論文誌の情報は階層構造化され、論文誌の年代、ナンバー、論文タイトル、著者、ページ番号等をXML構造で記述する。XML に書き込まれた情報は検索用インタフェースにより全文検索を可能とする。

(2)角度推定可能な文字認識方式

最初に、学習法と認識法について述べる。

① 学習法

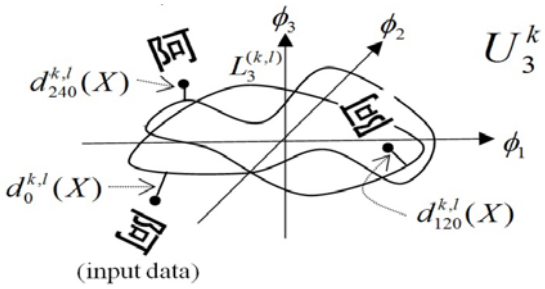
これまで英数字 62 文字種を対象に固有部分空間法を利用し、カテゴリ毎の固有空間を生成後学習文字を投影して、認識実験を行ってきた。本研究の認識対象は JIS 第一水準漢字、ひらがな、カタカナなど 3,133 文字種と極めて多く、さらに明朝体、ゴチック体を識別対象にした。すなわち、フォント・カテゴリ・回転角を出力することになる。全文字種の明朝体・ゴチック体文字の実画像を収集することは困難なので、プログラムにより自動生成する。固有空間の生成法はいくつか考えられるが、本研究では、出来るだけ小さいメモリ量とするため、同じ文字種の複数フォントの文字を使って固有空間を生成する。その部分空間に学習文字画像を投影すると、下図のような軌跡が得られた。これは文字種「亜」の明朝体、ゴチック体の回転軌跡である。同じ文字種は類似した軌跡を描くが、ゴチック文字は明朝文字より黒画素が多いので、第一固有軸の正方向に軌跡が位置する傾向にある。



②認識法

全 3133 文字種について固有空間を生成し、学習文字から軌跡を求める。これらの空間に未知文字を投影し、投影点と軌跡との距離を計算する。投影には認識結果の信頼性をあげるために二つの方法を提案している。一つは入力画像をそのまま投影する方法であり、他のひとつは入力文字を等角度回転してそれらを固有空間に投影する。下図は入力文字画像「阿」を 120 度ずつ 2 回回転して 3 つの文字画像を生成して固有空間に投影した例である。ここに生成文字数を R とする。下図の場合 R=3 である。この方法により偶発的な誤認識が少なくなり認識結果が改善される。認識実験はプログラムにより生成した文字画像を用いた場合と、カメラにより取得した実画像を用いた二つの場合で認識性能を検討する。

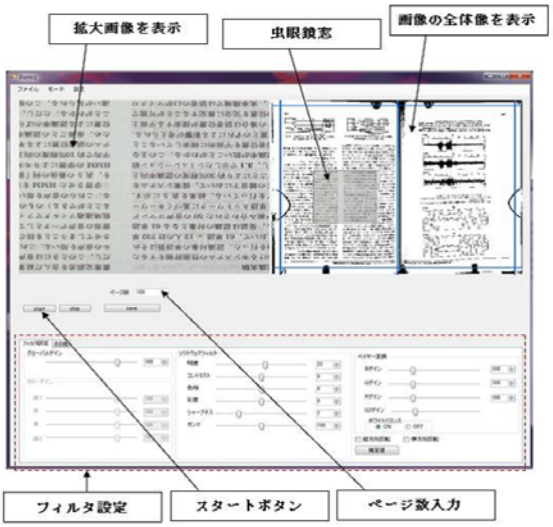
さらに、性能向上を目指すため、トップ 10 候補に対し、入力画像の投影点の最近傍点からその角度の文字画像を再構成して、入力画像と相関を取ることにより類似性を再評価する。

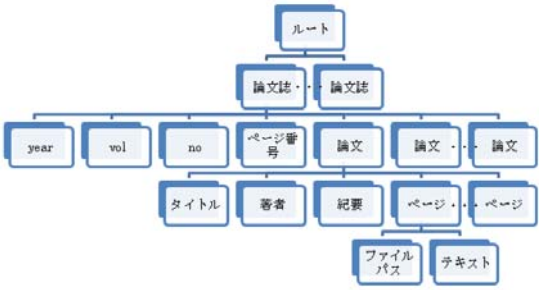


４．研究成果

(1) ページ捲り器と高解像度カメラを用いた文書処理方式の成果

ページめくり器と同期する高解像度カメラ制御用インタフェースを作成した(下図)。これにより、原画像、拡大画像の状況を観察しながら諸パラメータ制御が可能となった。また、適応的二値化、副作用の少ないノイズ除去法、再帰的領域分割など一連の文書解析・処理プログラムを作成した。さらに、簡易な論文データベース検索ソフトウェアを





作成した。キーワードにヒットした文字列をクリックすると、そのページの情報が得られる。XML で表現する論文誌群のデータ構造は上図のようである。

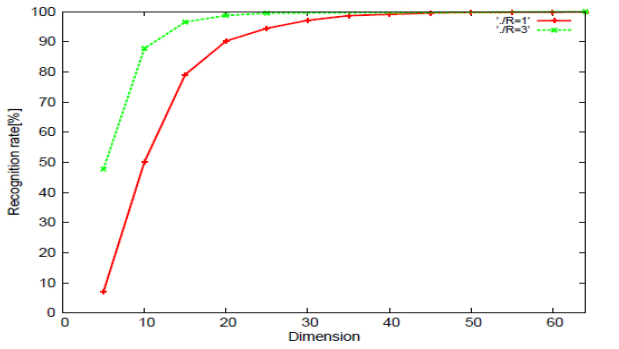
フリーOCR エンジンの Tesseract を用いた認識率は古い年代の紙面品質で 82.2%とあまり高くはないが、前述の一連の前処理をしなければ 39.2%と極めて低かった。

ページ画像列から先頭ページの検出精度を調べるため、電子情報通信学会論文誌 69 冊（1431 論文、13,156 ページ）を対象に先頭ページシンボルの検出実験を行った。シンボルは[論文]、[招待論文]、[ショートノート]、[サーベイ論文]、[研究速報]である。結果的に成功率は 84%となったが、失敗の原因としては、ページ捲り器によるページ境界の歪みに依るものが多かった。[論文]以外の先頭ページシンボルについては低い抽出結果となった。これは元々シンボルの数が少ないため 1 シンボルが占める比率が大きいことや、紙面の歪みによる場合が起因した。[研究速報]については、ページの端にあるため、見開きページ分割においてシンボルが切れてしまい、検出率を低下させた。本課題で生じた問題点を再度検討し改善する必要がある。

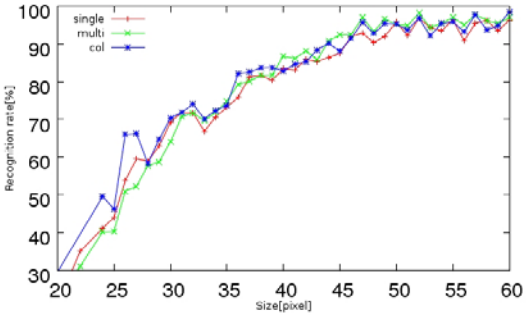
(2) 角度推定可能な文字認識方式の成果

①人工生成文字に対する次元と認識率の関係

下図は次元に対する認識率のグラフである。図中2つの曲線があるが、下の曲線は R=1の場合、上のグラフは R=3の場合である。日本語3133カテゴリを対象としているにも関わらず、高い認識率を示し、R=1の場合と比べ、R=3 の場合は低次元で高い認識率が得られることが実証された。

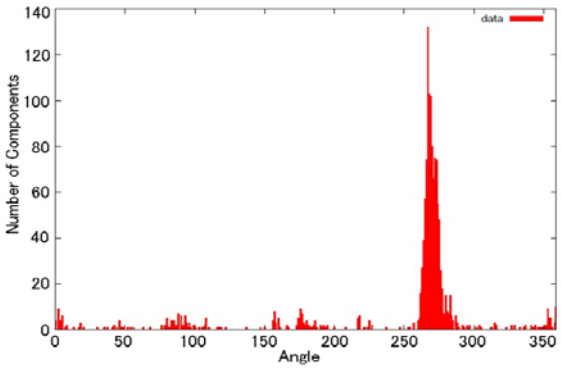


②カメラによる文字画像のサイズと認識率の関係



ここでは、カメラからの距離をいくつか設定して取得した文字画像について認識実験を行った。現実的な時間で実験を行うため、2フォント、75カテゴリを無作為に選んだ。故に実験に用いた文字数は150である。上図の横軸は文字画像の横幅(画素数)である。図中3つのグラフがあるが、青色がトップ10を対象とした再評価による認識率、緑色が R=3の認識率、赤色が R=1 の認識率である。文字数が少ないのでグラフが凸凹しているが、横幅が45画素以上の文字画像について、どの方法においても95%程の認識率が得られた。小さなサイズの文字では同じ文字種でも文字形状の相違が著しいためどの方法でも効果が表れなかった。

③実文書への適用



一般に実文書をレイアウト解析して文字認識すると、認識誤りがレイアウト解析に起因することがある。それを避けるために、前実験としてレイアウト解析をせず、直接連結成分に対し文字認識を適用し、統計的に文書の角度が推定できるかを検討した。ひとつの連結成分が一文字を形成しているとは限らないが、モロフォロジー演算をしてなるべく文字分離を減らしている。上図は原画像(新聞の一部)で左に約90度傾いている。この画像から連結成分を求め、本認識方式を適用した結果、連結成分の推定角度のヒストグラムが次のグラフである。角度の最頻値は271度であり、正しい結果であった。他の文書画像においても実験し、高い精度が得られることが明らかとなった。

今後、研究期間中に出来なかった課題を改善し、文書解析処理プログラムと角度推定可能な文字認識プログラムを統合する計画である。

５．主な発表論文等

〔学会発表〕（計　8 件）

① Yuki Tanaka,Hiroyuki Hase,Shogo Tokai,"Construction of an Academic Paper Database using a High Resolution Camera and Page Turning Machine", International Workshop on Advanced Image Technology (IWAIT2014), A2-P019,(2014).

② Yuta Baba,Hiroyuki Hase,Shogo Tokai, "Rotated Character Recognition for A Huge kinds of Categories", International Workshop on Advanced Image Technolog(IWAIT2014), P1-P058, (2014).

④ Yuta Baba,Hiroyuki Hase,Shogo Tokai, "Rotated Kanji Character Recognition", The 13th IAPR Conference on Machine Vision Applications (MVA 2013),(2013).

⑤ 馬場悠太,長谷博行,東海彰吾,”マルチフォント回転文字認識に関する研究”,電気関係学会北陸支部連合大会,F2-17,(2013).

⑥ 田中侑己,長谷博行,東海彰吾,"論文誌群から論文データベースの構築法",電気関係学会北陸支部連合大会,F2-27,(2013).

⑦ 佐藤詩織,長谷博行,東海彰吾,"日本語を対象とした回転文字認識",電気関係学会北陸支部連合大会,F-100,(2012).

⑧ 田中侑己,長谷博行,東海彰吾,"論文画像群からの先頭ページの判別",電気関係学会北陸支部連合大会,F-103,(2012).

③ Shukun Ning, Hiroyuki Hase, Shogo Tokai, ”Kanji Character Generation for Character Recognition”, 2011 Joint Conference of Hokuriku Chapters of Electrical Societies,F-56,(2011).

６．研究組織

(1)研究代表者

長谷　博行（HASE, Hiroyuki）

福井大学・大学院工学研究科・教授

研究者番号：　90142273