3C-7

黒画素連結成分の外接矩形による英文和文判定方式

福田 浩至　樋野 匡利　町田 哲夫
(日立製作所システム開発研究所)

1．はじめに

論文や新聞、雑誌等の印刷文書の構造を理解し、その内容を認識する文書画像処理技術の必要性、重要性が高まっている。具体的には、文書画像ファイルへの格納時のタイトルや著者名等のインデックス付け、機械翻訳システムへの初期入力や文書データベースの構築等への適応が求められている。本稿では、文書画像処理技術における、文字列の英文和文判定処理方式について報告する。

2．文書画像処理の概要

文書画像処理の手順の一例を図２．１に示す。以下各処理について説明する。

（１）メディア分離

文書画像から文字、図表、写真、画像領域等を分離、抽出する。メディア分離処理には、周辺分布法、ランレングス法、近接線密度法、連結成分法等の多くの方式が提案されている[1]。ここでは連結成分の外接矩形を用いて処理する。処理の方法は既に報告した[2]。

（２）文字列抽出

文字領域の中から文字列を抽出する。これに関しても多くの方式が提案[3]されているが、ここでは(1)のメディア分離処理で用いた連結成分の外接矩形を用いて、縦書き横書きを自動判定して文字列を抽出する[2]。

（３）文書構造理解

抽出された各領域の情報、文字領域内の文字列の情報をもとに、文書構造を理解する。即ち、ページ内の文字列や領域の配置に関する性質や文字の大きさ等の情報により、領域間のつながりや包含関係、さらにはタイトル、本文等、各領域の文書における意味を認識、理解する処理である。

この処理の対象は、明確な規則に従ってレイアウトされている文書に限定したもの[4]が多い。また、文書の多様性に対応する手段として、書式を予め定義しておき、それに基づいて処理を行う方法や、文書のレイアウトを記述する幾何的構造から論理的構造へ変換する方法等が提案されている。これに関しても、文字列の高さとその並びピッチを利用して、性質の類似する文字列を統合し、文書を構成する部分領域を抽出する方式を開発し、既に報告した[2]。

（４）文字切り出し

抽出された文字列から個別文字を切り出す。文字の並びの違いや、和文と英文の違い、隣接する文字が接触した場合の切りだしの有無等により、さまざまな報告[5]がある。和文と英文では、その性質が大きく異なることから、従来の手法では処理対象をどちらかに限定している。

本報告では、英文、和文両方を扱えるようにするために、まず両者の異なる特徴に着目して、文字列が英文であるか和文であるかを判定し、そのあとでそれぞれの文字列に応じた文字切り出しを行う方式を提案する。

（５）文字認識

切り出された文字を認識する処理[6]であるが、その対象により、手法と難しさが大きく異なる。認識対象は、手書き／活字の別や、英数字、片仮名、平仮名、漢字等の文字の種類によって分類される。

上述の各技術のついて、処理(1)、(2)と(3)の一部については報告済みであり、以下、３章では、処理(4)のうち、和文と英文の判定処理の詳細について述べる。

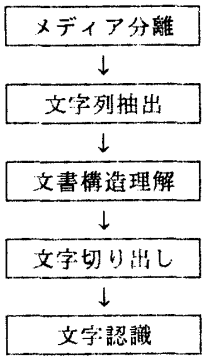

図２．１　文書画像処理の手順

Document Image Processing
Using Rectanglar Information
Kouji FUKUDA, Masatoshi HINO, Tetsuo MACHIDA
HITACHI Ltd.

３．和文英文判定処理

３．１　和文と英文の文字列の特徴

文書中の文字領域において、各文字を構成する黒画素の連結成分に着目して、漢字、片仮名、平仮名などからなる和文と、大文字、小文字のアルファベットからなる英文とを比較すると、その並び方において、次のような特徴を見出すことが出来る。

（ｉ）文字列垂直方向の特徴

英文の場合はベースとなる位置が存在して'a','c'のように上下につきでないもの、'b','d'のように上につきでるもの、'g','j'のように下につきでるもの等に分類できるが、和文の漢字、片仮名、平仮名にはそのような特徴はない。

（ii）文字列平行方向の特徴

英文では、単語と単語の間にスペース部分が存在するが、和文にはない。

（iii）文字を構成する矩形の特徴

英文は、単一の連結成分からなるものが多い（'i','j'以外全て）。一方、和文は、複数の連結成分からなるものが多い（特に漢字）。

従って、黒画素の連結成分の外接矩形（以下、連結成分矩形と略す）の上辺、下辺の位置や矩形の間隔にそれぞれの特徴が生じると考えられる。これらの特徴に基づき、文字列の和文英文判定を行う。

３．２　和文英文判定方式

文字列の外接矩形（以下、文字列矩形と略す）と連結成分矩形について、左上、右下の座標をそれぞれ(LXmin,LYmin),(LXmax,LYmax)と(RXmin,RYmin),(RXmax,RYmax)で表わす。

以下、ある文字列矩形Ｌとその中に含まれる連結成分矩形Ｒ(n)を用いて、判定方式を説明する。　３．１で述べた特徴（ｉ）、(iii)に着目した判定は、文字列に含まれる連結成分矩形の上辺と下辺のＹ座標、R(n)Ymin、R(n)Ymaxの Ｙ軸方向での頻度分布を求めることにより実行する。例えば、図３．１に示すように、文字列矩形をＹ軸方向に８分割し、R(n)Ymin、R(n)Ymaxの存在する領域の頻度分布を求める。この分布は図３．２に示すように、和文と英文で異なった特徴を示すと考えられる。領域ａ、ｈに高い頻度を示す文字列は和文、領域ａ、ｈの他に、その内側の領域ｂ、ｇにも高い頻度を示す文字列は英文と判定できる。

特徴(ii)に着目する方式では、まず図３．1に示た同じ文字列矩形内の隣接する連結成分矩形間の距離Ｄの分布を求める。Ｄの分布が、図３．３に示すように、２つの顕著のピークのある場合は英文、そうでない場合は和文と判定する。

４．おわりに

黒画素連結成分の特徴に着目し、文字列内の連結成分矩形を用いて、文字列の和文英文判定を行う方式について報告した。これにより、英文和文を判定した後、各々の特徴を利用した文字切り出し、文字認識を行うことができる。

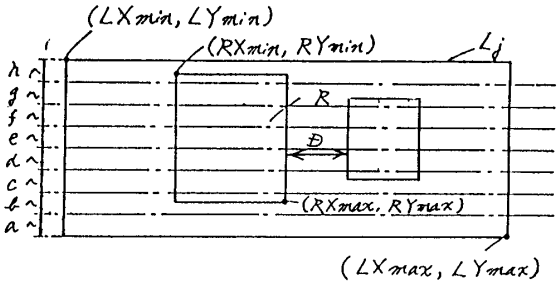

図３．１　連続成分矩形と文字列矩形の関係

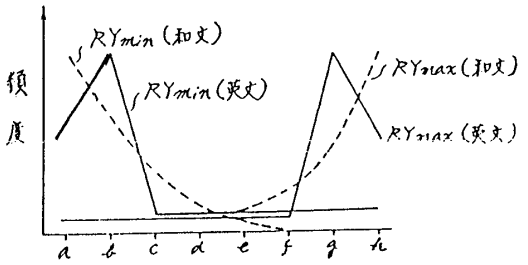

図３．２　上辺、下辺の位置の頻度分布

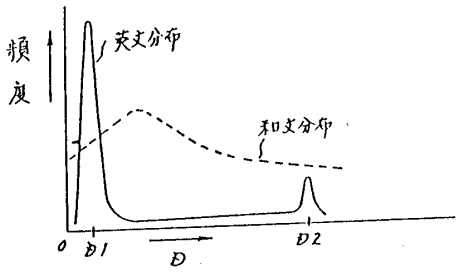

図３．３　距離Ｄの分布

参考文献

[1]秋山他:"周辺分布、線密度、外接矩形特徴を併用した文書画像の領域分割",信学論(D),Vol.J69-D,No8

[2]樋野他:"文書構成要素の分離抽出方式",第33回情処全大5Y-3

[3]辻他:"スプリット検出法に基づく頁画像の構造解析",信学技法,PRL85-17

[4]山田他:"記事の形状に着目した英文新聞の領域分割",第26回情処全大2B-4

[5]中村他:"横書き日本語文書における個別文字の切り出し法",信学論(D),Vol.J68-D,No11

[6]津雲他:"文字認識技術の最近の動向",情処研資,IE88-5