3C-6

# 不定ピッチ文字列を含む印刷文書を対象とした文字切り出し手法の検討

佐藤道弘　木田博巳
NTTデータ通信株式会社

## 1．はじめに

日本文中に英単語が混在する印刷文書において文字切出しを行う場合、ピッチ情報のみをもとに切出しを行うと、文字ピッチ不定のアルファベットの部分で、誤った結果を得る可能性が高い。本稿では、各々の黒画素連結領域について、それが全角文字であるか、その一部であるか、又はアルファベット文字であるかを推定する方法、および、その結果を文字列のレイアウトルールにより修正して文字切出しを行う手法を提案し、評価実験の結果を述べる。

## 2．処理対象文書

対象とする文書は、日本文主体の印刷・横書き文書で、中に含まれているアルファベット文字には、ピッチ不定のプロポーショナル文字を含むものである。今回は黒画素連結領域の情報のみを使って処理を行ったため、接触のある文字の強制切出しについては検討を行っていない。

## 3．処理手順

処理手順の概略を機能ブロック図の形で図1に示す。

### (1) 前処理部

前処理部では、イメージ・リーダ（400dpi）から入力された文書について各行の切り分けを行った後、各行について、黒画素連結領域（8連結）の抽出を行い、その位置、大きさを得る。さらに、同じ行内にある領域について、垂直方向に重なる部分を持つものを一つに統合する。（以下この統合された黒画素連結領域を統合矩形と呼ぶ。）

### (2) スコア算出部

スコア算出部では、前処理で得られた統合矩形の情報をもとに各々の矩形が表1に示すカテゴリーのうちどれに属する可能性が高いかをスコアとして算出する。
スコア算出のために、まず、各統合矩形について、大きさ（高さ幅）、行内での上下方向の位置の偏り、プロポーション、左右の矩形との距離等の特徴量を求め、これと辞書との対比によりスコアを決定する。

### (3) 判定部

スコア算出部で求められたスコアは、個々の統合矩形の性質を反映したものである。判定部では、さらに、文字列としての特徴を文字切出しに反映するために、各矩形のレイアウト上の特徴をルール化し、このルールを反復的に適用してスコア操作を行い、最終的な切出し結果を求める。

表1　統合矩形のカテゴリ分類

| No | カテゴリ | | 説明（例） |
|---|---|---|---|
| ① | 全角文字 | 単独 | 単独の統合矩形で文字を形成（亜　愛） |
| ② | | 左 | 偏の部分（何のイ、外のタ） |
| ③ | | 右 | 旁の部分（何の可、外のト） |
| ④ | 句読点 | | （。　、　，　．） |
| ⑤ | アルファベット | タイプ1 | （b d f h i k l t） |
| ⑥ | | タイプ2 | （g j p q y） |
| ⑦ | | タイプ3 | （a c e m n o r s u v w x z） |

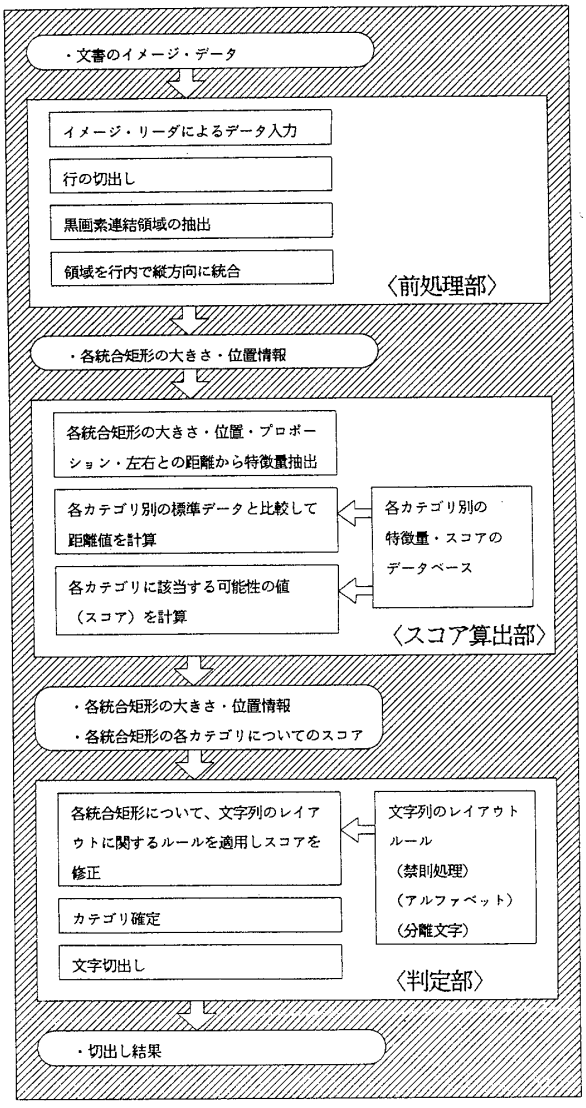


図1・処理手順

---

Character Extraction Method for Japanese Documents Including Alphabetic Character Strings
Michihiro SATOH,　Hiromi KIDA
NTT DATA COMMUNICATIONS SYSTEMS CORPORATION

## 4．評価実験

（1）人間の識別能力

本手法の限界を見極める目的で、統合矩形の枠情報のみを被験者に提示し、人間の切出し能力を評価した。その結果、未訓練の被験者で９０%～９８．５%の正解率、訓練後の被験者で最高１００%の正解率を得た。

（2）スコア算出のための特徴評価

各統合矩形の大きさ（高さ・幅）、行内での上下方向の位置の偏り、プロポーション、左右の矩形との距離等カテゴリ分類に有効と思われる特徴量をヒューリスティックに選択し、それぞれの分布をとって特徴評価を行った。

図２は、特徴量の１例であり、行の高さに対する統合矩形の上下のすき間の高さの比率を示している。統合矩形５１１３個についての分布をアルファベット文字・タイプ３とそれ以外に分けて示したのが図３である。アルファベット文字・タイプ３では、その値は、０．５２付近に集中し、他ではそれよりも小さい値をとるものが大半である。ヒストグラムの各区間について、アルファベット文字・タイプ３の占める比率をとると（図３－ｃ）、この比率についての平均・分散は、それぞれ、0.52, 0.0023 となり、図２に示す特徴量がアルファベット文字・タイプ３のスコア計算に有効であることがわかる。

このような特徴量評価を、各カテゴリについて複数個行い、その結果に基づき、スコア計算のための特徴を選択した。

（3）スコア操作の効果

各矩形についてスコアが最大となるカテゴリを推定結果とした場合、約８０%の正解率を得た。ここで誤りとなったものは２文字以上続いたアルファベットについて、全角分離文字の左側と右側としてのスコアが高く計算された場合が多かった。そこで、アルファベット文字列については、その文字列としての特徴（文字の並びかた）をルール化し、このルールによって、スコアを操作する。

ルールの例を次に示す。

当該矩形について、
（1）．スコアが全般にしきい値より低く
（2）．アルファベットのスコアが高い矩形に挟まれている
場合、その矩形のアルファベットのスコアを増加させる。

このようなルールを適用し、スコアを操作することによって、各矩形のレイアウト上の特徴を反映させることが出来る。

レイアウト上の特徴を取入れたスコア操作の後、文字切出し判定を行った結果、正解率として９７%を得た。（表2）

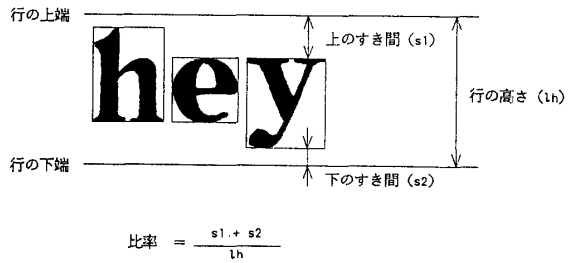


図２・抽出する特徴量の例

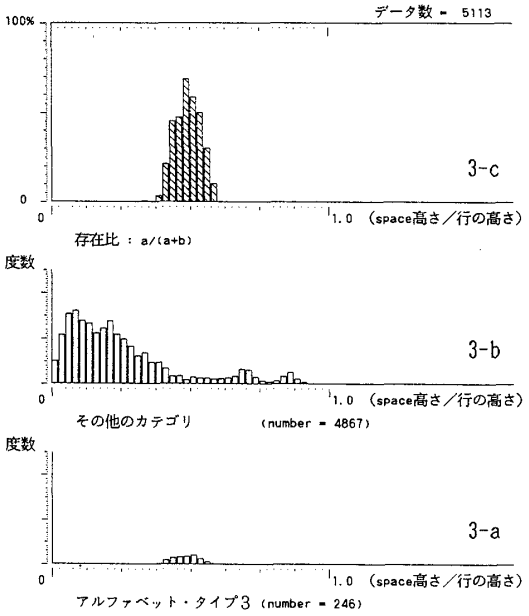


図３・特徴量の分布

表２・実験結果

| | 矩形を単位としたカテゴリ分類の正解率） | 文字を単位とした切り出し率 |
|---|---|---|
| スコア算出の後、最大スコアを採用した場合 | 77.3% | 93.8% |
| 判定部でスコアを修正した後の最大スコアを採用 | 85.4% | 96.5% |

## 5．むすび

黒画素連結特徴の抽出、各統合矩形の文字カテゴリに該当する可能性（スコア）の算出、アルファベット文字列等のレイアウト上の特徴によるスコアの補正によって、精度の高い文字切出し手法を構成出来ることを示した。

今回の実験では、統計的性質の異なる文書を一緒に扱い、スコア算出用辞書を作成したため、必ずしもカテゴリ分類に有効な距離値が得られなかった面もある。スコア算出部では比較的単純な距離値の計算によってスコアを決定出来るため、複数の文字フォント別にスコアのデータベースを保持すれば、現在不足な部分も含めて更に精度は向上すると考えられる。また判定部においてはルールは簡単な条件によって構成されており、追加・改善は容易である。スコアは最後まで保持されるため、ＯＣＲによる文字読み取りの際、切出し方に優先順位を設定し、フィードバック処理をすることも可能であると考えられる。