I-027

多重仮説検定を用いた割書・振り分け行に対応した文字行抽出方式

An Extraction Method of Complex-Structured Text Lines Based on Multiple Hypothesis Testing.

高橋 寿一†，古川 直広†，今泉 敦博‡，藤尾 正和†，永崎 健†，渡辺 成‡，酒匂 裕†

Toshikazu Takahashi, Naohiro Furukawa, Atsuhiro Imaizumi, Masakazu Fujio, Takeshi Nagasaki, Shigeru Watanabe, Hiroshi Sako

1. まえがき

文書画像理解において，文字行抽出技術は従来から研究が行われているが，帳票処理の対象拡大に伴い，多様な文字行，例えば図 1 に示す割書行や振り分け行にも対応する必要がある．

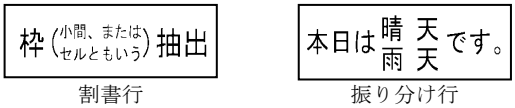
割書行・振り分け行に対応した文字行抽出技術の一つとして[1]の方式がある．これは連結成分の隣接関係をネットワークで表し，そのネットワークの部分構造が分岐・環状・変化している箇所から切断点を抽出し，切断点に基づいて矛盾の無い文字行を全て抽出する．しかしこの方式では，通常の文字行において，「嶺岩」のように部首が上下に分かれる文字行では，切断点が得られずに上下に分かれた文字行しか抽出できないと考えられる．

上記課題を解決するために，我々は多重仮説検定に基づく方式を提案する．この方式は，まず，連結成分の重なりから一次文字候補を生成し，そして縦方向に並んでいる文字候補同士を結合したものを二次文字候補として生成する．上下の部首を結合した候補も二次文字候補の中にできる．次に，これらの文字候補から直線性や近傍性等を検定して文字候補同士を結合して文字行を生成する．これにより複数の文字行候補を抽出することで，割書行や振り分け行，通常行の候補も含まれる．最後に，抽出された各文字行候補の検定は，文字認識結果と単語照合により検定する[2].

本論文では，帳票内の横書きの文字行を対象に，文字行の候補を出力するまでの方式を検討し，実験結果より方式の効果を確認する．

2．割書行と振り分け行の特徴

本論文で対象とする文字行の特徴を図 1 で示す．割書行は，1 行の間に 2 行（または複数行）を書き入れている行である．このような書式は主に割注として用いられている．また振り分け行は，1 行中でその行と同じような大きさの文字を使って 2 行（または複数行）に行を分けて書いたものである．主に選択肢や並列の意味を持たせるのに用いられている．



割書行　　振り分け行

図 1：割書行と振り分け行の記載例

† （株）日立製作所中央研究所

‡ （株）日立製作所情報機器事業部

3. 提案方式の概要

3.1 処理手順

本方式の処理手順を図 2 に示す．入力は連結成分の外接矩形座標である．ここで，原画像は傾いている場合があるため，まず連結成分の外接矩形座標を回転させて傾きを補正する．次に傾き補正後の連結成分の外接矩形において，ノイズ等の過小成分は除去する．以降の処理は傾き補正後のノイズを除去した外接矩形座標で行う．次にこの外接矩形座標から多重仮説検定に基づき文字候補を生成する．その文字候補を結合して文字行候補を生成する．最後に包含関係等から文字行候補を絞込み，文字行候補を出力する．図 2 の太枠の処理について，次節より詳細に説明する．

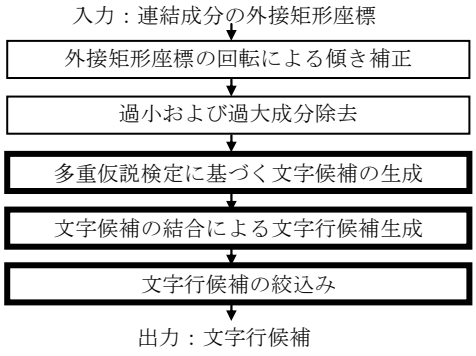


図 2：本方式の処理手順

3.2 多重仮説検定に基づく文字候補の生成

文字行を抽出する前段階として，多重仮説検定に基づき文字候補を生成する．

まず，一次文字候補を生成する．図 3 で示すように，外接矩形同士が交差・接触・包含の関係であるならそれらを結合する．結合できる成分が無くなるまでこの処理を繰り返す．これを一次文字候補とする．

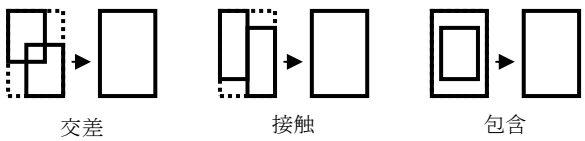


図 3：一次文字候補の生成

次に，一次文字候補を基に二次文字候補を生成する．一次文字候補の段階では，例えば「嶺」という文字では「山」と「領」が分離してしまう．そのため，縦方向に並んでいる一次文字候補同士「山」と「領」を結合し，それを新たに二次文字候補「嶺」とする．この場合も，二次文字候補と縦方向に並んでいる一次文字候補がある場合，更に結合を繰り返して新たな文字候補を作成する．

ただし，この結合はある閾値の範囲で繰り返す．以上の処理を行うことで図 4 に示すように，1 つの文字からでも「山」「領」「嶺」という異なる文字候補が生成できる．

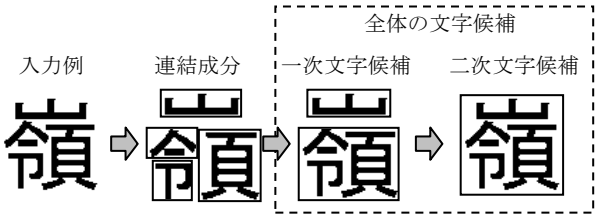


図 4：多重仮説による文字候補生成例

3.3 文字候補の結合による文字行候補生成

次に，文字候補を基に文字行候補を生成する．ある 1 つの文字候補が文字行を生成していくための核になり，別の文字候補との関係が以下の条件下であれば結合して，それを新たな核とする．

(1) 横方向に重なり部分がある
(2) 距離が近い
(3) 高さの差が小さい
(4) 高さの差が大きいが，ハイフンやコンマ，中点，小文字と思われる配置にある
(5) 直線的に並んでいる
(6) 2 つの間に他の文字候補がない

この操作を繰り返し，結合できる文字候補が無くなるまで行う．このような処理により様々な高さの文字行候補が生成できる．ここで，一度結合した文字候補は他の文字行候補生成時には結合対象外にする．ただし，上記条件の(4)の場合は，他の文字行候補生成時には上記条件の(3)である可能性があるため，結合対象にする．

以上の処理により，図 5 に示すように，1 つの行からそれぞれの文字候補の大きさに基づいた文字行候補が生成できる．また，結合されなかった文字候補は 1 文字の文字行である可能性もあるため，この文字候補も文字行候補として扱う．

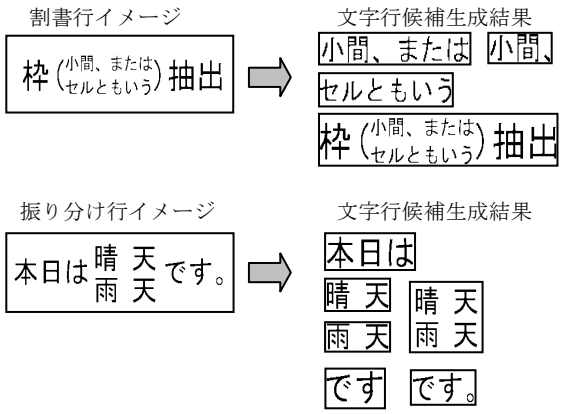


図 5：文字候補の結合による文字行候補生成

3.4 文字行候補の絞込み

最後に，以下のいずれかの条件により文字行候補を除外し，文字行候補を絞込んで出力する．

(1) 文字行候補の生成時に結合しなかった文字候補が，文字行候補と包含関係である．
(2) 十分に長い文字行候補において，文字行を形成する連結成分の射影により切れ目がある．

4. 実験結果

本対象である割書行・振り分け行を用いて実験した．抽出結果は目視により確認した．文字行抽出結果を表 1 に示す．ここで割書行の正解とは，書き入れられている複数行が個々の行として抽出でき，かつ全体を 1 行として抽出できている場合である．また振り分け行については，振り分けられている複数行が個々の文字行として抽出でき，かつ振り分け前後が個々の行として抽出できている場合とした．

対象サンプル 93 行に対して 86 行が正解であった．失敗した 7 件については，行間接触による文字候補生成の失敗 3 件，文字間が広過ぎることによる文字候補結合の失敗 3 件，カスレによる文字候補結合の失敗 1 件であった．

表 1：文字行抽出結果

対象サンプル	正　解	失　敗
93	86	7

また通常の文字行を用いて，「嶺岩」のような上下に分離できる複数の文字から構成される文字行において，正しい文字行が抽出できるか実験した．例えば「嶺岩」という文字行に対して，「嶺岩」という文字行が候補に含まれていれば正解とし，「山山」と「領石」という文字行しか候補の中になければ失敗とする．100 サンプルを用い，全て正解の文字行が候補に含まれていることを目視により確認した．これらの実験により，本方式の有効性が確認できた．

5. まとめ

本方式を用いることで，割書行・振り分け行の文字行抽出が可能となった．この方式の特徴は，連結成分を基に多重仮説検定に基づき文字候補を生成し，生成した文字候補同士を直線性や近傍性等により結合して文字行を求める．

本方式により，割書行や振り分け行からの文字行が抽出できるようになり，また上下に分離できる複数の文字で構成される文字行においても，正しい文字行が候補の中に含まれていることで，本方式の有効性を確認した．

本稿では横書きの文字行のみを対象としたが，帳票には縦書きの文字行も含まれている．今後の課題としては，横書きと縦書きの両方の文字行に対応することである．

参考文献

[1] 宇陀明弘,他, “隣接関係ネットワークに基づく文字列抽出”, 電子情報通信学会総合大会, D-460, p.248(1996)
[2] 古川直広,他, “星座認識による帳票識別方式”, 信学技報, PRMU2001-125, pp85-92(2001)