I-85

帳票画像からの下線抽出の一手法

An Extraction Method of Underlines in Form Images

嶋 好博† 新庄 広† 丸川 勝美† 中島 和樹†
Yoshihiro Shima Hiroshi Shinjo Katsumi Marukawa Kazuki Nakashima

1．まえがき

文字読取りの対象である帳票には実線、点線、破線、鎖線などの罫線が印刷されている。罫線の抽出や識別並びに除去処理は文字読取りのために不可欠である[1][2][8]。下線は下線付きの文字行と通常の文字行とを区別する重要な情報である。文字切出しや文字識別処理に当該下線は障害となるため、下線を抽出し、下線と文字行とを分離する必要がある。本研究は、色々な罫線のうちで、特に、文字行の下側に印刷されている下線を抽出する手法を提案する。

下線を抽出する処理過程は一般に、(a)罫線の抽出処理、(b)文字行の抽出処理、(c)罫線と文字行との配置判定処理の三つの処理からなる。この内、(c)下線の配置判定処理は、大別すると、（1）文字行方向の投影パタンを算出し、線分位置の近傍にある黒画素の分布の偏りから、下線と所定の配置関係にある文字行を判定する手法（画素投影分布法）[5]、（2）線分の端点（始点、終点）座標を求め、線分端点から所定位置にある文字行を判定する手法（二次元座標値法）[3][6]がある。（1）の画素投影分布法は郵便の宛名行の書式判定に提案されており[5]、黒画素を文字行方向に投影し、この投影分布のピークを下線位置としている。しかしながら、線分の長さが短い下線の場合、投影分布の偏り等の特徴が顕著でなく、下線と文字行の配置関係の判定が困難である。また、帳票が傾いている場合、分布のピークを判定できないという問題がある。（2）の二次元座標値法は、グラフ構造の解析[3]や小切手の記載項目抽出のために提案されている[6]。小切手には、予め記入用のガイドライン（罫線）が印刷されており、ガイドライン（罫線）を抽出し、罫線から所定位置にある領域内の文字行を抽出している。この（2）の手法を単純に本研究の対象帳票に適用できない。何故なら、帳票には一文字幅程度の短い下線があり、文字の要素であるストロークと短い罫線との識別が困難で偽りの罫線が抽出されるという問題があるためである。

提案する新手法は、（2）の二次元座標値法に属しており、矩形の配置関係を利用した構造解析[3]を下線抽出のために拡張した手法である。先ず、所定値より長い黒いランを連結し罫線の候補群を求める[4]。また、罫線抽出処理と独立し、黒画素の連結成分を融合して文字行を抽出する[7]。そして、抽出した文字行の位置情報をもとに、これら罫線の候補群の内、文字行の下部に配置する罫線を下線として選択する。

2．下線の特徴と下線抽出の技術課題

2.1 下線の特徴

帳票に印刷されている下線には次のような性質がある。

① 方向はほぼ水平であるが6度以内の傾きがある。
②原画像では下線は文字行から一定値下側に離れているが、縮小した低解像度画像では下線と文字行とが接触している場合がある。
③多くの文字行では文字行の長さと下線の長さはほぼ等しいが、文字行の一部分に下線が付いている場合がある。
④一つの文字行に下線が複数個付いている。
⑤一文字幅の文字行に短い下線がある。
⑥下線の線幅は一定である。

図1は文字行と下線とが接触している例である。図2は一文字幅の文字行に下線が付加されている例である。

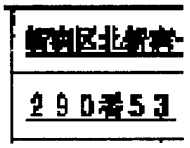


図1 文字行と接触する下線

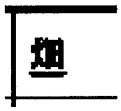


図2 一文字幅の下線

2.2 下線抽出の技術課題

下線と文字行との配置関係は、通常は文字行の下側に文字行とほぼ同じ長さの下線が印刷されている。また、一つの文字行に下線が複数本付いている場合もある。さらに、長さが短く1文字幅の下線の場合もある。下線を抽出するための課題を列挙する。

①偽罫線と下線の識別, ②傾いた文字行と下線の配置関係の判定, ③文字行と接触した下線の検出, ④短い（一文字幅の）下線の抽出, ⑤一つの文字行にある複数下線の抽出

3．下線の抽出処理

3.1 処理の概要

帳票読み取りの処理フローは、先ず、スキャナから帳票画像を採取する。次いで、画像を4分の1に縮小する。これは、罫線抽出、文字行抽出の高速化のためである。縮小画像に対して、罫線の抽出を行う。罫線抽出では所定値より長い黒ランを連結し、この細長い連結成分を罫線として取り出す[4][7]。また、抽出した罫線から傾き角を検出する。さらに、黒画素の塊を融合して文字行を抽出する[7]。そして、下線抽出に移り、求めた罫線群と文字行群からそれらの配置関係を判定して下線と対応する文字行を選択する。次いで、文字行座標を基に、原画像から文字行を切り出す。切り出した文字行画像に対して、下線を除去する。このとき、先に抽出した下線の端点座標を基に下線上の黒いランを除去する。下線除去後の文字行画像に対して、文字を切り出し、各文字パターンに対して文字識別を行う。

下線抽出は、偽罫線の発生を防止するため、次節に述べるように、文字行の長さによって、複数文字を含む文字行の下線抽出と一文字のみ含む一文字幅文字行の下線抽出とに分けて処理している。

† （株）日立製作所中央研究所,
Central Research Lab., HITACHI, Ltd.

3.2 複数文字を含む文字行の下線抽出

予め、文字行の座標と傾き角を文字行抽出処理及び罫線抽出処理において求める。文字行は図３に示すように、文字行を取り囲む外接矩形の頂点座標であらわす。文字行の傾き角を利用して、文字行の上辺と下辺の端点座標を算出する。罫線は始点と終点の座標であらわす。注目する文字行の近辺には、下線だけでなく、偽りの罫線や他の文字行に付加された下線が多数存在する。このため、文字行の下側において、文字行の上辺ならびに下辺から所定距離にある探索範囲を設け、対応する罫線を選択する。

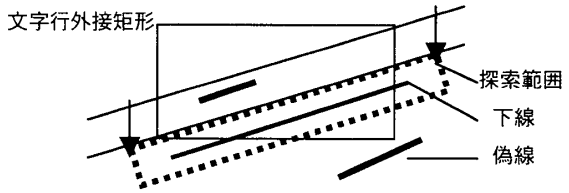


図３ 文字行の下線探索の説明図

3.3 一文字幅文字行の下線抽出

一文字幅文字行の下線抽出では、先ず、文字行の中から所定値以下の幅と所定値以下の高さをもつ文字行を探索する。そして、この寸法条件を満たす文字行について、その下側に探索範囲を設ける。黒画素の連結成分を探索し、所定のサイズをもつ連結成分の外接矩形が当該文字行の下側の探索範囲に存在すれば、その連結成分は下線であると判定する。さらに、下線が接触している文字行に対応するため、一文字幅文字行の中から所定値以下の幅と所定値以上の高さをもつ文字行を探索する。この一文字幅文字行の下線抽出の前提は、文字の高さと幅が既知という条件である。

4．下線抽出の実験

4.1 下線抽出実験システム

下線抽出の実験システムは画像入力装置、帳票認識装置からなる。画像入力装置はモノクロイメージスキャナ(400dpi,フラットベッド)を使用し、画像ファイルに一旦原画像を格納した。帳票認識装置としては、ワークステーション(CPU クロック 125MHz)を用いた。入力した画像は処理の高速化のため、1/4 に縮小した 100dpi 相当の画像を用いて、罫線抽出、文字行抽出を行い、これらの結果を利用して下線を抽出する。実験プログラムはＣ言語で作成した。

4.2 下線抽出実験結果

図４は下線抽出の処理結果を示している。矩形で示す文字行の下側に線分で抽出した下線を表示している。図５は一文字幅の下線の抽出結果の例である。図６は下線除去後の文字切り出し結果の例である。

少量の帳票（９７１枚）画像サンプルで下線抽出の精度を求めた。その内訳は、下線付きサンプル２３９枚、下線なしサンプル７３２枚である。見逃しと虚報のサンプル数を求めた。見逃しサンプルとは、下線が付いているにもかかわらず下線を検出しなかったサンプルである。また、虚報サンプルとは、元々下線のない個所で誤って下線を検出したサンプルである。下線付きサンプルの内、見逃しが２枚、虚報が２枚あったが、これらは手書きノイズや汚れが原因であり、実用上問題はない。

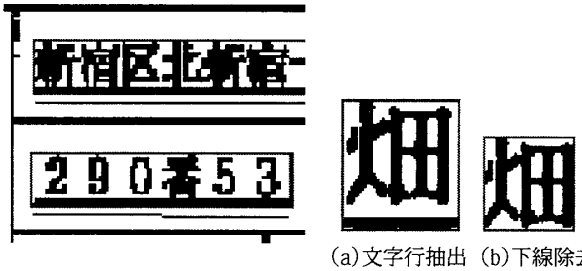
5．むすび

帳票画像から下線を抽出する手法を考案し、少量の画像サンプルで下線抽出実験を行い、手法の有効性を確認した。

下線抽出の今後の課題としては、以下の項目がある。

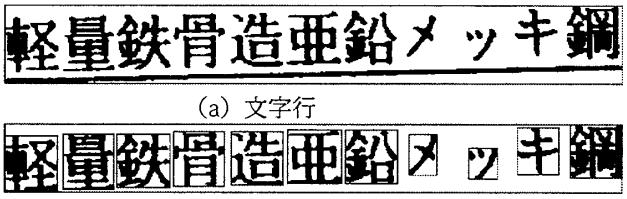
(a)手書き文字行と下線との接触、交差対応

(b)下線の上側及び下側の文字行との接触と強制分離



(a)文字行抽出 (b)下線除去

図４ 接触下線の抽出結果　図５ 一文字幅の下線抽出結果



(a) 文字行

(b)下線除去と文字切り出し

図６ 下線抽出と下線除去の結果

6．参考文献

[1] R.G.Casey and D.R.Ferguson :" Intelligent Forms Processing", IBM Systems Journal, Vol.29, NO.3, pp.435-450, 1990

[2] Yasuaki Nakano, Yoshihiro Shima, Hiromichi Fujisawa, Jun' ichi Higashino, Masaaki Fujinawa:" An Algorithm for the Skew Normalization of Document Image", Proc. 10th Int. Conf. Pattern Recognition, pp.8-13, June 1990

[3] Masashi Koga , Tatsuya Murakami , Yoshihiro Shima , Hiromichi Fujisawa : "An Extraction Method of Search Indexes for Graph Image Retrieval ", MVA' 92 IAPR Workshop on Machine Vision Applications, pp.163-166, Dec. 1992

[4] 古賀昌史，中島和樹，丸川勝美，嶋好博，藤澤浩道："棒状図形の傾き検出の ラン符号による高速化の一手法"，情報処理学会第４６回（平成５年前期）全国大会，8C-6, pp.2-219~2-220(1993 年)

[5] Noboru Nakajima, Tetsuo Tsuchiya, Takeshi Kamimura, and Keiji Yamada:" Analysis of Address Layout on Japanese Handwritten Mail -A Hierarchical Process of Hypothesis Verification-", Proc. 13th Int. Conf. Pattern Recognition , pp.726-731, Oct. 1993

[6] Ke Liu, Ching Y. Suen, and Christine Nadal:" Automatic Extraction of Items from Cheque Images for Payment Recognition", Proc. 13th Int. Conf. on Pattern Recognition , pp.798-802, August 1996

[7] H.Shinjo, E.Hadano, K.Marukawa, Y.Shima and H.Sako :" A Recursive Analysis for Form Cell Recognition", Proc. ICDAR' 01, pp.694-698, Sep. 2001

[8] 嶋好博，新庄広，丸川勝美，中島和樹："帳票画像からの点線抽出の一手法"，2002 年電子情報通信学会総合大会，D-12-60, p.236(2002 年)