ＦＡＸ送信票からの文書構成要素の抽出

4G－3

小西英樹　中島崇博　本田康弘　鈴木秀智　樫尾次郎
三重大学　工学部

1　はじめに

現在、文字認識の研究として様々な研究が行われており、文字が切り出されていることを条件に活字文字、手書き文字のいずれもかなり高い認識率が得られている。そこで、受信したＦＡＸ送信票から宛先等を抽出、認識し、電子メールで各個人に配送するシステムの構築を試みている。

文書構成要素として枠線、罫線といった線分や写真、表といった図形、文字列などが文書中には含まれており、なかでも線分は領域の分割を明確にするなど、文書構造に関する重要な情報を与える。また除去することにより、空白帯を用いた領域分割や文字（列）抽出が行いやすくなる。

本稿では、大局的な性質を示す周辺分布特徴と局所的な性質を示す黒画素連結成分の外接矩形特徴を用いて、文書構成要素である線分と文字列を抽出する手法について検討したので報告する。

2　処理の流れ

本稿では、ワープロ等で宛先等の記入欄があらかじめ作られており、手書きで宛先等を書き加えてあるＦＡＸ送信票を200dpiで受信したもの（図１）を対象にしている。以下に処理の流れを示す。

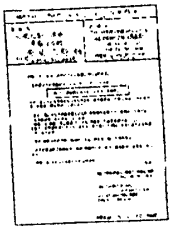

図１　原画像

2．1　前処理

前処理として、LPP法[1]を用いた傾き補正、８近傍の平均値フィルタを用いた平滑化を行い、ラベリングを行い、連結成分の画素数が８dots以下の連結成分をノイズとして除去する。

Extraction of Document Component from FAX cover page.
Hideki Konishi, Takahiro Nakashima, Yasuhiro Honda
Hidetomo Suzuki, Jiro Kashio
Faculty of Engineering, Mie University

2．2　線分要素の抽出／除去（１）

前処理において連結成分の画素数が小さい連結成分をノイズと判定した。反対に、画素数（a_1：入力画像による）の大きい連結成分は文字以外の領域ここでは線分要素候補領域と考えられる（図２）。

各領域に対して水平、垂直両方向の周辺分布を求め、線分の抽出を行う。横線の場合、垂直方向の周辺分布の値が大きいところは縦線が存在するので、平均値以上となる列を除去した横線候補求める（図３）。

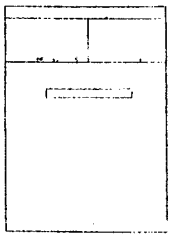

図２　線分要素候補

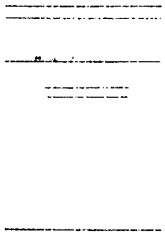

図３　横線候補

縦方向成分を除去することにより明確な水平方向の周辺分布を得ることができ、これを用いて横線を抽出、除去する。縦線の場合も同様に行う。

2．3　領域分割

枠線等の除去を行ったことにより空白帯を用いた領域分割が可能となったので、階層的領域分割法[2]を用いて性質の異なる領域へ大分割する。ＦＡＸ送信票など文字数が少ない文書の場合、文字間隔が開いた文字列の影響で空白帯が現れ、主空白帯と誤って判定し、誤分割を行うことがある。そこで、領域の中央付近に存在する空白帯のみ主空白帯と判定することで誤った領域分割をなくしている。

大分割された領域に対して、先に抽出された線分及び周辺分布の空白帯を用いて小分割を行う。各領域において、線分が領域の高さ（幅）の７０％以上を占める線分をフィールドセパレータとし領域分割

を行う。分割された領域が最小文字サイズC_{min}以上であれば文字列候補領域に、C_{min}未満であれば線分及びノイズ領域と判定する。

2．4　線分要素の抽出／除去（2）

連結成分の画素数を用いて抽出できなかった点線や短い線分を連結成分の外接矩形特徴を用いて抽出除去を行う。外接矩形が次の条件を満たすものを点線候補とする。

矩形高さ ≦ C_{min} かつ 黒画素占有率 ≧ 0.5

点線候補に対して、次の統合処理を行う。外接矩形の垂直方向の中心が、統合相手の外接矩形の高さの範囲にあり、統合を行おうとしている矩形間において、矩形幅の比が3以下の時、次の条件の満たす矩形の統合を行う。

矩形間隔 ≦ 2×矩形幅 又は 統合回数 ≧ 4

統合回数が5以上の矩形を点線として抽出、除去する。

独立した線分は横縦比が2以上で、最大文字サイズをC_{max}とした時

矩形幅 ≧ C_{max} かつ 矩形高さ ≦ C_{min}

を満たす矩形を線分として抽出、除去を行う。

文字と接触した線分の場合、各外接矩形の水平、垂直両方向の周辺分布を求め、水平方向の周辺分布の値が平均値以下かつ、垂直方向の周辺分布の値が平均値以上の画素を除去したものを横線候補とし（図4）、明確になった水平方向の周辺分布から横線を抽出、除去する。

図4　接触線分の抽出

抽出した線分を境界線として、外接矩形の垂直方向の中心が境界線より上に存在する矩形の集合と下に存在する矩形の集合に領域分割する。

2．5　文字列抽出

各文字列候補領域の高さが推定文字サイズの2.2倍未満のとき文字列とし、2.2倍以上のとき、複数の文字列が存在すると判定し文字列の抽出を行う。文字サイズは、外接矩形の高さの平均から求めるが、最小文字サイズより小さい外接矩形は、推定文字サイズを小さくするという悪影響を及ぼすので、あらかじめ矩形相互の重なりが、矩形面積の0.1以上となる矩形の統合を行っておき[3]、統合後の矩形の高さがC_{min}以上の矩形の平均を推定文字サイズとする。

高さがC_{min}以上の矩形を垂直方向の中心座標で面積を投影値として水平方向に投影し、0が連続する部分の長さLがC_{min}以上の範囲で、水平方向の周辺分布が最小となる位置で文字列を分割、抽出する（図5）。

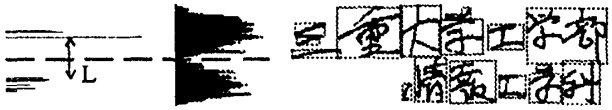

図5　文字列の抽出

3　実験結果

これまでに述べた手法により、線分および文字列抽出実験を１６枚のＦＡＸ送信票に対して行ったところ、２１２本の線分のうち２０９本の線分を抽出し、文字列は３６３個のうち３４４個抽出できた。

線分抽出では、あらかじめ候補を絞り込むことで誤抽出を抑えることができた。文字列抽出での誤抽出はいずれも手書き文字を含む文字列であり、分割位置は求まるが、上下の文字列間で文字の接触や誤った外接矩形の統合が起こり、上下の文字列を1つの文字列に抽出する誤りが見られた。外接矩形を強制的に切断して、文字列抽出を行う必要がある。

4　終わりに

周辺分布特徴と外接矩形特徴を用いて線分及び文字列を抽出する手法を提案した。今後は、抽出された結果を用いて、文字切り出し及び文字認識を行う予定である。また、一般的な文書に対して実験を行い、汎用性についても検討を行う予定である。

参考文献

[1]　秋山，増田：書式指定情報によらない紙面構成要素抽出法,信学論(D),J-66-D,1,pp.111-118 (1983)

[2]　朴，海老名，伊藤：汎用的な文書画像の階層的領域分割と識別法,信学論(D-II),J-75-D-II,2,pp.246-256(1992)

[3]　馬場口，塚本，相原：手書き日本文字列からの文字切り出しの基礎的考察,信学論(D),J-69-D,12,pp.2123-2131(1986)