文字列抽出型周辺分布法による文書の傾き検出

6W-1

秋山 照雄　　宮原 末治
NTTヒューマンインタフェース研究所

1．はじめに

画像として入力された文書を処理するためには、傾きを予め検出し正規化しておく必要がある。文書の傾きを検出する方法として、文字の外接矩形を用いるものなど種々の方法[1]～[3]が報告されているが、ここでは特徴の抽出が容易な周辺分布を用いた手法について述べる。この手法は、局所的な周辺分布の中から文字列に対応する部分を選択し、それを用いて傾きを検出するもので、これまでに提案してきた手法[4]では十分な精度が得られなかった図表を含む文書や、段ごとに文字列の位置が異なる文書に対しても、高精度な傾き検出が可能となる。

2．傾き検出アルゴリズム

従来、提案してきた局所的な周辺分布を用いる手法(LPP法[4])は、文書の傾きを周辺分布の相関演算によって検出する位相差検出型のものであった。ここで提案する文字列抽出型の周辺分布法は、小領域内で求めた周辺分布の中で、文字列に対応する部分のみを用いて傾きの検出を行なう。処理のステップを以下に示す。

(1) 文書画像を高さhのたんざく状の小領域に分割し(図1(a))、各々の周辺分布P(x)(図1(b))を求める。

(2) 周辺分布の投影成分(P(x)が一定値以上となる範囲、図1(b))を求める。投影成分Cは左端及び右端のx座標$C(x_1)$、$C(x_2)$で表現する。中点をC(xm)とする。

(3) 投影成分の幅($C(x_2)-C(x_1)$)のヒストグラムの最頻値より文書の文字列の幅を推定する。

(4) 互いに隣接した領域の投影成分の中で、その幅が(3)で推定した文字列幅の許容範囲にあり、さらに上下方向に見て互いに重なり合っているもの
$Cn(x_2)\geqq Cm(x_1)$ かつ $Cn(x_1)\leqq Cm(x_2)$
に同一のラベル①を付与する(図2(a))。これにより、文書中の図や写真などを取り除くことができる。

(5) 同一のラベルを持つ投影成分の中で、上下方向の重なりの小さいもの、すなわち一定の条件
$Cn(x_1)\geqq Cm(x_1)-\alpha$ かつ $Cn(x_2)\leqq Cm(x_2)+\alpha$
を満たさないものには新たなラベル②を付与する(図2(b))。αは領域の高さhと入力時に許容される最大の傾きから設定する。

(6) 同一のラベルを付与された各々の投影成分の傾きを最小自乗近似によって求める。投影成分の傾きはそのまま文字列の傾きとなる。全体の平均を求めて傾きの仮抽出結果とする。

(7) 同一のラベルを持つ投影成分の中点が一直線上に乗らないもの、すなわち、一定の条件
$Cn(xm)\geqq Cm(xm)+offset-\beta$ かつ
$Cn(xm)\leqq Cm(xm)+offset+\beta$
を満たさないものをさらに分離し、新たなラベル③を付与する(図2(c))。この処理によって図3に示すような段ごとに位置が変化する文字列も分離可能となる。(6)と同様、最小自乗近似によって文書の傾きを求めて最終結果とする。offsetの値は(6)で得られた傾きの仮抽出結果から求める。βは許容誤差である。

3．実験結果と考察

実験には16本/mmのファクシミリで入力したデータを4本/mmに変換したものを用いた。データの種類を表1に示す。データ1,3,5は段ごとに文字列位置の変化がある。

表1　実験に用いたデータ

番号	内容	方向	図表	文字列位置変化
1	朝日新聞記事	縦	あり	あり
2	ｻｲｴﾝｽ記事	横	あり	なし
3	朝日新聞記事	縦	あり	あり
4	特許明細書	横	なし	なし
5	英雑誌記事	横	あり	あり

(1)傾き正規化文書画像の作成

入力した画像に本手法を反復適用することにより傾きを補正した正規化文書画像を作成した。正規化文書画像は画像中に参照ラインを引くことにより傾きがないかどうかを確認した。データ3について-5度傾いたもの(図3(a))と正規化されたもの(図3(b))を示す。

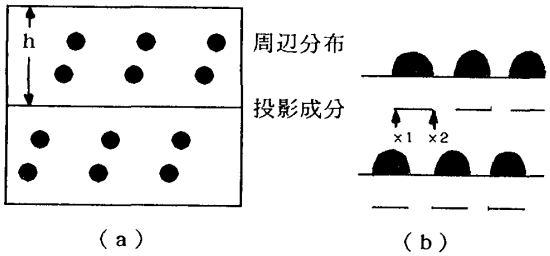


図1　たんざく状の小領域と周辺分布の投影成分

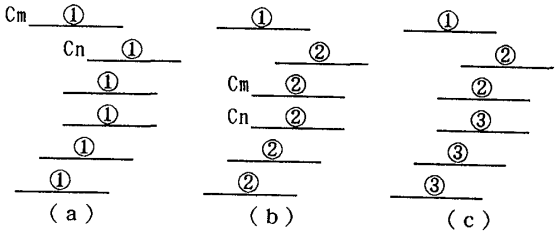


図2　投影成分のラベリング

Document Skew Detection
based on Text Line Extraction
Teruo AKIYAMA, Sueharu MIYAHARA
NTT Human Interface Laboratories

(2)傾き検出範囲と精度

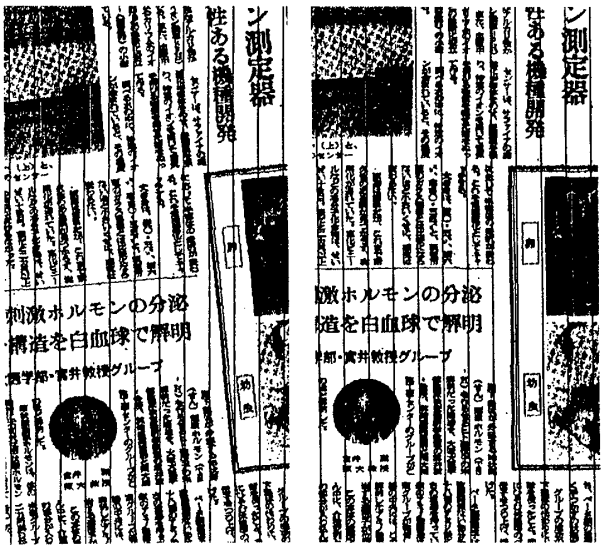
各データについて傾きの検出範囲を求めたものを図4に示す。A4サイズの新聞の入力を考えた場合、隣接する文字列の周辺分布を分離可能とするために許容される最大の傾きは約0.3度である。そこで、誤差が0.15度以内で求められる範囲も同時に示した(図4、斜線部分)。文書を入力するときの傾きが通常は5度以内であることを考慮すると、実質的に問題のない精度で傾きの検出が行なわれていることがわかる。データ3は図や写真の領域が広く、段ごとに文字列の位置が異なるため(図3)、傾き検出精度が低くなっている。データ3の結果については(4)でさらに検討する。

(3)傾きの検出限界

新聞の社説欄から見出し文字列を取り除いたデータを作成し、傾きの検出限界について調べた。文字列間の距離をd、分割した小領域の高さをhとすると、傾きの検出限界予測値はarctan(d/h)となる。実験で求めた傾き検出範囲と、予測値の関係を表2に示す。この図から、予測される範囲とほぼ同じ範囲での傾きの検出が行なわれていることがわかる。なお、このデータの文字列の幅は約2.8mm、文字列間の距離は約1.6mmであった。

表2　傾き検出範囲予測値と実測値

領域の高さh	予測値(度)	実測値(度)
16	±21	−22〜+20
32	±11	−10〜+9
64	±5	−4〜+7
128	±3	−2〜+1



(a)　(b)

図3　参照ラインによる正規化画像の確認

(4)入力文書の傾きの大きさと検出精度

データ3の正規化画像を回転させた時の回転角度と検出角度の関係を図5に示す。誤差が0の場合には各点は45度の直線上にのる。この図から、データ3のように条件の悪いデータについても比較的安定に傾きの抽出ができていることがわかる。なお、このデータはLPP法[4]による傾きの検出が不可能だったものである。

4．まとめ

文書画像を複数個の領域に分割して求めた局所的な周辺分布の中から、文字列を反映している部分を選択し、それを基に文書の傾きを検出する文字列抽出型周辺分布法を提案し、その有効性を実験によって確認した。今後は、文字列が少ない図を主体とした文書の傾き検出が課題である。

文献

(1)射手園ほか、昭63信学全大D-475
(2)中野ほか、信学論(D), J69-D, 11, pp1833-4
(3)長谷ほか、信学論(D), J67-D, 9, pp1044-51
(4)秋山ほか、信学論(D), J66-D, 1, pp111-8

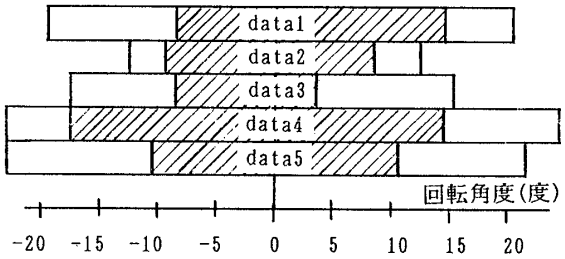


図4　傾きの検出範囲と精度(斜線内誤差0.15度以下)

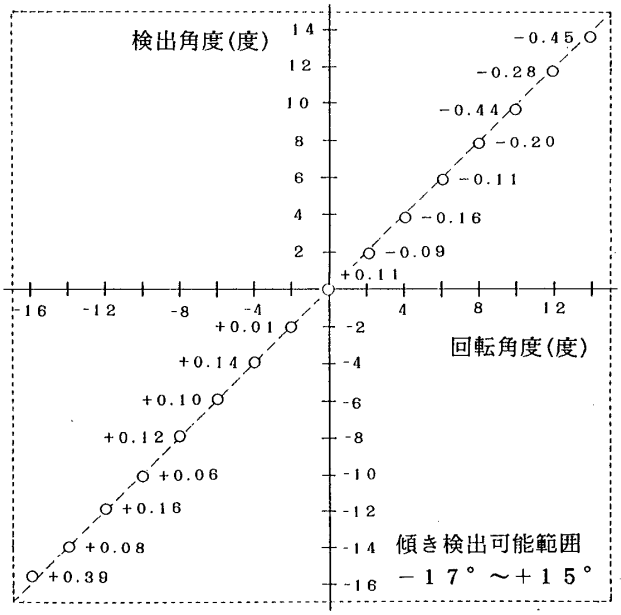


図5　回転角度と検出角度の関係(ﾃﾞｰﾀ3)