D-11-99

線画の混在する画像におけるテキスト領域の抽出に関する検討

A Study on Text Area Detection in Binary Images Containing Line Drawings

河村 圭　　山本 勇樹　　渡辺 裕
Kei KAWAMURA　Yuki YAMAMOTO　Hiroshi WATANABE
早稲田大学大学院 国際情報通信研究科
Graduate School of Global Information and Telecommunication Studies, Waseda University.

1　はじめに

テキスト検索やインデキシング技術の向上により電子化テキストの有用性が広く認識され，既存印刷文書（活字）の電子化需要が高まっている．そこで，紙媒体をスキャナによりディジタル化し，画像中から文字を抽出，認識する技術が必須である．

線画や文字が混在する画像では，文字のピッチや大きさなどの事前情報から文字領域を抽出するのが困難である．そこで，従来から研究されている手書き文字列から文字を切り出す方式に着目する．これらの手法では，外接矩形の面積や縦横比などの形状的特徴を文字の判定に用いる．さらに，文字認識結果や単語照合結果を文字切り出しにフィードバックする方式も検討されている．しかし，形状的な特徴に基づく切り出しがより正確になれば，フィードバック情報の確度が高まり，性能がより一層向上する．このため，文字の形状的特徴のみに基づく文字切り出しの高精度化は，重要な課題である．

本稿では，既存の文字認識を適用するために，画像から文字領域を抽出することを目的とする．また，コンテンツとして需要の高いマンガを主な対象画像とする．

2　従来手法

文字領域の抽出に先立ち行を抽出する手法がある．文字と思わしきもの（外接矩形）を文字行に垂直な直線に射影し，その分布密度を用いて行を抽出する．

後藤ら [1] は，罫線やノイズの混入を考慮しない単純な外接矩形が誤抽出の原因であることを示した．そこで，文書画像中の区分直線状の要素を文字行と仮定して抽出することにより，これを解決している．文書構造に関する知識を必要とせず，画像のゆがみにも耐性がある．

本手法は必要なしきい値が多く，特に文字サイズに対するロバスト性が十分でない．また，罫線以外の要因に対する考慮が十分でなく，線画が混在すると誤検出が増加する．

3　提案手法

文字の形状的特徴を用いて，罫線や線画の誤検出防ぐ．さらに様々な文字サイズに対応するため，小さい文字から大きい文字まで順番に文字サイズ S を仮定して処理を進める．以下に S における抽出手順を述べる．

■ 外接矩形の取得　連結黒画素から外接矩形を得る．矩形の長辺が S より長い場合は，その矩形を破棄する．

■ 文字列中心の取得　注目する矩形の周辺を探索して，分布する矩形から文字列密度を計算する．得られた密度をしきい値を用いて行が存在するかを判定し，存在する場合には文字列中心の座標を求める．密度は文字の面積に対して 2 割の小ささまで許容する．文字列に垂直方向へ $S \times 2/3$，水平方向に $S \times 4/3$ を探索範囲とする．

■ 分離した文字の統合　得られた文字列中心に対して，文字列と垂直方向に分離している矩形を統合し，より大きな矩形を得る．

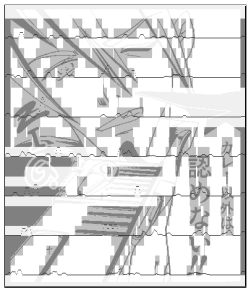

図 1　従来手法の結果

図 2　提案手法の結果

■ 文字列の取得　得られた文字列中心により，文字列と水平方向に分離している矩形を統合して，文字列を取得する．分離した文字を統合する条件は，矩形の中心同士の距離が文字列に垂直方向に $S \times 2/3$，水平方向に $S \times 4/3$ 以下で，かつ重なりが 0.5 以上とした [2].

■ 文字列群の統合　得られた文字列を，文字列方向と垂直に拡張して文字列群を取得する．統合した矩形の周囲に余白が存在する場合，孤立文字として抽出する．

本手法は，画像の局所的な特徴を利用しているため，画像のゆがみや文字の位置に対してロバスト性が高い．

4　実験と考察

対象画像の代表例であるマンガを 200dpi でスキャンし入力画像とする．しきい値などを文献通りにした従来手法の矩形と行密度を重ね合わせた結果を図 1 に，提案手法により得られる文字の矩形，行密度，行中心を重ね合わせた結果を図 2 に示す．

従来手法では，線画も含めて行密度に反映しているため，後処理の行の判断をしきい値処理により行えない．一方，提案手法では長い線画をあらかじめ分離可能であるため，文字列候補となる矩形の精度が高い．さらに，文字列の中心付近を各文字で捉えられており，文字列を抽出可能である．

5　おわりに

本稿では，線画と文字が混在する画像から文字領域を抽出する手法を提案した．提案手法は，文字の形状的特徴と画像の局所的な特徴を利用しているため，線画や画像のゆがみや文字の位置に対してロバスト性が高い．実験により，文字列中心が精度良く抽出でき，提案手法の有効性を確認した．

謝辞

この研究は，財団法人大川情報通信基金研究助成による．

参考文献

[1] 後藤，阿曽，“文字行の局所的な直線性を利用した頑健・高速な文字行抽出法，” 信学論 D–II，vol. J78–D–II，no. 3，pp. 465–473，Mar. 1995.
[2] 仲林，北村，河岡，“あいまい用語検索を用いた高速枠なし手書き文字列読み取り方式，” 信学論 D–II，vol. J77–D–II，no. 11，pp. 1528–1537，Nov. 1991.