カラー文書画像画像からの文字列抽出†

4N－04

長谷川 史裕§
（株）リコー　ソフトウェア研究所¶

1.はじめに

文書画像から文字領域を特定・抽出することはOCR処理にとって不可欠である。近年、カラープリンターの普及などによりカラー文書を扱う機会が急速に増している。従来、カラー画像に対しては、一旦二値画像に変換してから文字領域の特定を行う場合がほとんどであった[1]が、色の異なる文字がある場合は適切な二値化ができず、対応が難しい。また、明度／色情報を用いた方法では色クラスタリングを用いる方法[2]や濃淡画像から細長い平坦なテクスチャを抽出する方法[3]があるが、小さな文字の抽出が難しく、解像度を上げて対処しようとしても網点に対応するのが難しくなる。

本稿では、カラー文書画像から色情報を用いて文字列を抽出する手法を述べ、精度の定量評価と処理時間についても示す。

2.文字列抽出手法

(1)圧縮画像生成　原画像の隣接する数画素四方を一画素にまとめた圧縮画像を生成する。圧縮画像の画素値は原画像数画素四方の最も明度の低い画素値、または高い画素値を用いる。以後、前者を最暗画像、後者を最明画像と呼ぶ。圧縮により、ストロークの細い小さい文字、背景に網目状のドットパターンがある場合でも対処がしやすくなり、処理の高速化も期待できる。まず最暗画像で処理を行って背景に対し暗い文字を抽出する一連の処理を行った後、この過程に戻って今度は最明画像で明るい文字を抽出するという繰り返し処理を行う。（図1参照）

(2)連結成分抽出　圧縮画像において、互いに隣接する画素で色の近いもの同士を連結成分として抽出する。

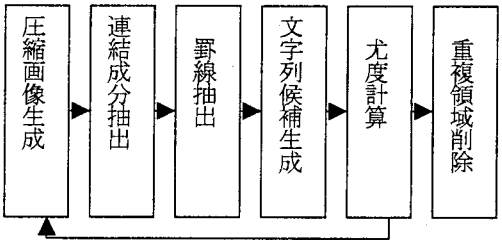

図1　処理フロー

(3)罫線抽出　連結成分を構成する画素で、長く連続して存在するものだけからなる連結成分を再構成し、これを罫線として抽出する。

(4)文字列候補生成　連結成分の外接矩形（以下単に矩形と書く）のサイズを考慮し、互いに近くにある矩形同士をグループ化して文字列候補を生成する。その際、色が異なるもの、間に罫線があるものは同一グループとしない。

(5)尤度計算　文字列候補のらしさ（尤度）を計算する。複数の特徴量から計算する。特徴量は文字列内画像のエッジ強度、文字列の構成連結成分の大きさなどを用いた。

n種ある特徴量のうちのひとつをiとし、実際の文字列での値の平均値をEi、分散をVi、尤度を求めたい文字列候補での値をFiとした場合の尤度Lは

---

†Extraction of Character Line from Color Document Images

§Fumihiro HASEGAWA
¶Software Research Center, Ricoh Co. Ltd.
1-1-17,Koishikawa,Bunkyo-ku,Tokyo,112,Japan

$$L=\sum_{i}^{n}(F_i-E_i)^2/V_i$$

で定義する。Lが所定値より大きい候補は文字列でないとみなして削除する。

(6)**重複領域削除**　(1)でも述べたように暗文字、明文字を(1)～(5)を繰り返して抽出した後、文字列同士が重なっている場合は尤度に応じて文字らしくないほうを削除する。

3. 処理結果

本手法の性能評価のため、実際のカラー文書画像を用いて文字列抽出率の測定を行った。比較のため、適応二値化を行ってから二値画像に対して処理を行う方法の文字列抽出結果も示す。

抽出対象画像総数 208、文字列総数 24386、200dpi、A4 サイズの 24bit カラー画像に対しての文字列抽出率は表 1 のようになる。

手法	抽出率(学習原稿)	同未学習[%]
二値化法	74.83	68.75
本手法	90.83	78.53

表 1　文字列抽出精度

本手法では図 2(上)のように、同一背景上に異なる色の文字があっても抽出できる。さらに文字に影がついている場合でも尤度処理により文字のほうを正しく抽出できる。また、図 2(下)のようなストロークが細くぼやけた小さい文字に対しても（高さ 2 ミリ弱）抽出が行える。大きな文字では 2.5 センチまで対処可能である。一方、処理時間は CPU が Athlon1.2GHz の PC で測定したところ、平均 5.0 秒であった。

4. 今後の課題

文字列が縦横混在している場合への対処を行う。また、実用化のために更なる処理時間の短縮と精度向上を図る。

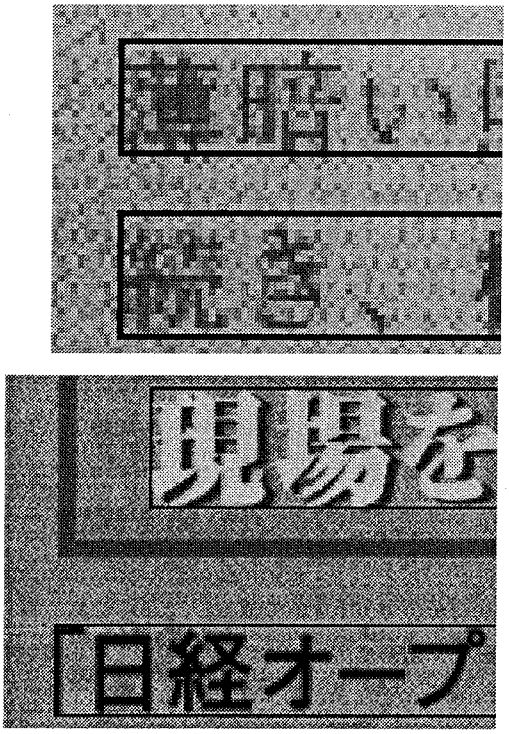

図2　処理例。黒線は処理結果を上書きしたもので幅１画素。原画像はカラー。

参考文献

[1]松尾賢一他「適応しきい値法を用いた情景画像からの看板文字列領域の抽出」,信学論,Vol.J80-D-II,No.6,pp.1617-1626,1997

[2]長谷博行他「カラー文書画像中の文字領域抽出を目的とした色分割についての検討」,信学論,VolJ83-D-II,No.5, pp.1294-1304,2000

[3]顧力栩他「表紙画像からの文字領域抽出方法」,信学論, Vol.J80-D-II,No.10,pp.2696-2704, 1997