カラー文書画像からの写真領域抽出手法

3M－3

久保亮二*　仙田修司**　美濃導彦**　池田克夫**
*三田工業株式会社　**京都大学工学部

1　はじめに

紙面に印刷された文字やパタンのデータを計算機を用いて自動的に分類，整理をするためには，文字領域と写真領域をあらかじめ分離しておくことが望ましい．文書画像からの写真領域抽出に関する研究の多くは，文書画像は背景の色が白，文字パタンの色が黒，写真領域は中間調と仮定しており，中間調領域の抽出や，隣接画素との濃度差を用いて写真領域を抽出している．

しかし，カラー文書画像を濃淡化した場合，必ずしもこの仮定は成り立たない．そこで，本稿では，積極的に色情報を利用して写真領域を抽出する手法を提案する．

2　カラー文書画像を扱う上での問題点

単色刷りの文書では，用紙の色がそのまま背景（地）となっていることが多いのに対して，色彩刷りの文書では，“チント”と呼ばれる単一中間色の部分的な背景が用いられることが多い．このチント領域上には，文字パタンや写真領域が重畳して置かれる．カラー文書画像の処理を行う場合には，チント領域から，文字パタンや写真領域を分離する必要がある．しかし，チント領域は写真領域と同様に網点で印刷されているため，濃淡化して局所領域でみた場合，写真領域とチント領域との区別をつけるのは極めて困難である．

人間がカラーの文書画像を見たとき，まず背景色とは異なる領域（前景物）に注目し，その前景物が文字パタンか写真領域かチント領域か判断しているのではないかと考えた．

そこで本手法では，まず背景色を取り除くことによって前景物を抽出し，抽出された前景物が文字パタンか写真領域かチント領域か判断する．チント領域ならば再び領域上の前景物を抽出するというように，再帰的に処理を行うことにより，文書画像上の全ての写真領域を抽出する．この手法を用いれば，関連性が大きいと思われる同一チント領域上の文字やパタンのデータを，あらかじめ分類することも可能であり，画像データの分類，整理にも有効である．

Extraction of Photographic Region from Color Document Images
*Ryoji KUBO
**Shuji SENDA, **Michihiko MINOH, **Katsuo IKEDA
*MITA Industrial Co., Ltd.
**Faculty of Engineering, Kyoto University

なお，カラー文書画像の背景色及び各文字パタンの色は単色であると仮定する．

3　写真領域抽出手法

写真領域抽出手法のアルゴリズムの構成図を図1に示す．

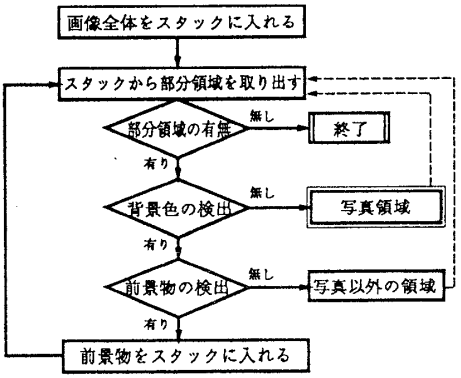

図 1: 写真領域抽出の手順

まず，画像全体を部分領域としてスタックに入れる．なお，カラー文書画像は網点で印刷されている領域が多いため，あらかじめ画像全体を平滑化しておく．

次に，スタックから一つ部分領域を取り出して，背景色の検出を行う．背景色を特定するために，カラー画像の色分解を行う．色分解は，均等知覚色空間においてクラスタを検出することにより行う．均等知覚色空間には，完全拡散面の3刺激値を250としたCIE1976($L^*a^*b^*$)空間を用いた．色分解を行うための色差d（$L^*a^*b^*$空間におけるユークリッド距離の2乗）は，人間の目で見て明らかに異なる色が判別できる程度に，適当に決めればよい．

色分解を行った後，各クラスタに属する画素の，画像上での分布を，水平垂直方向に射影することによって調べる．この結果，水平方向，垂直方向のほぼ全てのライン（全ライン数に対してP_p%以上）に一つのクラスタに属する画素が確認できた場合，そのクラスタが背景色に対応しているとする．背景色と判断されたクラスタが複数存在する場合は，その中で最も画素数が多いクラスタが背景色に対応しているとする．これにより背景色が得られなかった場合，部分領域は写真領域とする．この操作を，スタックに部分領域が残っ

ている間繰り返す.

背景色が検出された部分領域に対しては，前景物の検出を行う．まず，背景色画素を “0”，背景色以外の画素を “1” として，“1” を値に持つ画素の連結成分を取り出す．一般に，文書画像上の写真領域は，ある程度以上の大きさをもっているので，連結画素数がある閾値 T_f 以下の成分は，ノイズ，小さな文字パタンと判断し，閾値 T_f 以上の連結画素成分を前景物とする．また写真領域の縦横比は，極度に大きくならないため縦横比が 1:r 以上で，かつ短辺の長さが閾値 l_s より短い前景物には，写真領域が存在しないとしてあらかじめ取り除く．

以上の結果得られた前景物を，部分領域として再びスタックに入れる．前景物が抽出できなかった部分領域には，写真領域がなかったとして，その部分領域に対する処理を終える．

スタックに部分領域がなくなるまで，以上の処理を繰り返し，画像上の全ての写真領域を抽出する．

4 実験および結果

実験に使用した画像を図 2 に示す．400dpi でスキャナで読み込んだ 24bit/pixel のカラー文書画像を平滑化し，水平垂直方向をそれぞれ 1/4 に縮小して用いた．平滑化は，平均値フィルタを用い，縮小には，単純間引きを用いた．この画像に対して，色差 d を 50 としてクラスタリングを行うと 84 のクラスタが検出された．画素数が多い順に五つのクラスタをとり，P_p を 95%として背景色の検出を行ったところ，代表色（クラスタの平均色）が (R, G, B)=(203, 206, 221) のクラスタが，背景色と判断された．

背景色が検出されたため，原画像に対して前景物を抽出を行う．連結画素数の閾値 T_f を 800 画素として前景物を抽出すると，11 の前景物が抽出できた．得られた前景物を図 3 に示す．ここで縦横比の r を 3，短辺の長さの閾値 l_s を 40 画素として，写真領域が含まれてない前景物を除去すると，(c)(d)(e)(h)(i) が前景物として抽出された．抽出された前景物は再びスタックに入れる．

スタックから取り出した (d)(e) には，背景色が検出できなかったので，写真領域と判断した．(c)(h)(i) は背景色（(h)(i) は表の罫線部分が背景色とされた）が検出されたため，再び前景物の抽出を行った．その結果 (c) のみに四つの前景物が確認された．得られた前景物を図 4 に示す．これらを再びスタックに入れ，同様の処理を行うと，(c-2) が写真領域として抽出された．

以上の実験の結果，図 3 の (d)(e)，図 4 の (c-2) を写真領域として抽出することができた．

図 2: 原画像

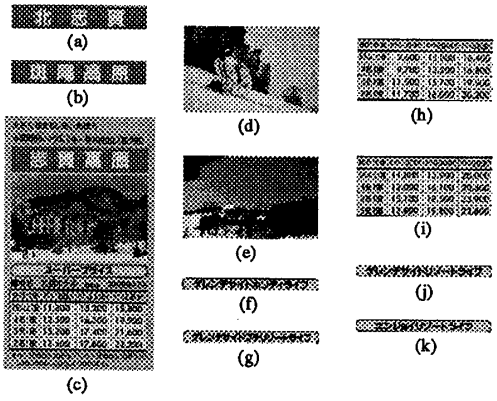

図 3: 原画像から抽出された前景物

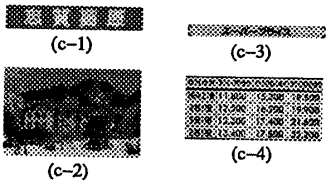

図 4: 図 3(c) から抽出された前景物

5 おわりに

カラー文書画像に対する，新たな写真領域抽出手法を提案した．本手法の特徴は，背景色との色差に注目して前景物を抽出し，得られた前景物の背景色を検出することにより写真領域を特定することを，再帰的に行うところにある．これにより，抽出された写真領域がどの前景物上にあるのか特定できる．このことは，紙面データの分類，整理という点においても有効である．

参考文献

[1] 富永: “カラー画像の色分解と分割,” 情処論, vol.31, no.11, pp. 1589-1598, 1990.

[2] 仙田, 美濃, 池田: “カラー画像からの文字パタン抽出法,” 第 47 回情処全大, 2-113, 1993.

[3] 長尾: “画像認識論,” コロナ社, 1983.