I-006

帳票画像におけるカラー背景からの文字パターン抽出の実験的検討

Experimental Study of Character Pattern Extraction from Color Form Images

久保田裕紀† 渡辺啓太† 吉田陽介† 嶋好博† 大矢博史†
Hiroki Kubota Keita Watanabe Yousuke Yoshida Yoshihiro Shima Hiroshi Ohya

1. はじめに

近年の情報技術の著しい発達に伴い、情報ツールにも大きな変化が見られてきた。それによって、紙書類からデジタルメディアに変換する必要性も増加してきている。また、デジタルカメラなどの発達と普及を受け、デジタル画像はより身近なものになっている。

本研究は、それに伴って出現した「紙書類をデジタルメディアとして認識させる」というアプローチの中の一つであり、カラー画像の２値化に関する研究である。画像を２値化して文字等を抽出する。２値化する方法としてはよく知られている大津法[1]を使用する。カラー画像の三原色成分[2]の平均値を抽出して大津法を適用させる方法を採用した。いろいろなカラー背景を持つ帳票から文字パターンを抽出する実験を行い、課題を明らかにする。

2. 目的

本研究では、帳票画像、すなわち紙書類等を一般的なスキャナ等で読み取り、その画像から文字背景を取り除き、必要な文字パターンのみを取り出すというのを目標としている。サンプル画像としてカラー背景中に文字列が存在している帳票を使っている。帳票のＩＤ部を読み取るため、文字列のみを抽出するという目的である。帳票に対しては種類・ID 部による選別、整理などのニーズがある。現在、これらの作業やデータの入力は人の手作業に頼る部分が多く、時間がかかり効率がよくない。これらの作業が自動になれば格段に効率のアップが見込められる、と考えられる。

3. カラー背景からの文字パターン抽出

3.1 ２値化の概要

２値化とは、各画素の明るさを一定の基準値により、黒色と白色の２色に変換する処理のことである。また「一定の基準値」を「閾値」と呼ぶ。通常、画像の各画素は、0～255 の RGB 値を持っており、この RGB 値の平均値が各画素における明るさになる。この明るさを使って２値化を行う。基準とする閾値の値によって、閾値処理後の画像は異なり、ノイズやかすれが発生する。

本研究で採用した明度は色の明るさを示したもので、以下の式で表すことができる。

$$K=(R+G+B)/3$$

ただし、R, G, B は対象画素の各色成分値、K は明度である。上式で示す明度 K を対象として、大津法により閾値を自動的に設定する。

3.2 大津法

閾値を自動で決定するために、本研究では「大津の方法」を使用する[1]。大津の方法とは、「ヒストグラムを２つの領域に分けるときに、それぞれの平均からの誤差が最小になるようにする」というものである。本研究は大津法により自動で閾値を決定する機能も持たせている。

3.3 三原色成分の平均値に基づく２値化方法

まずカラー画像から注目する画素に対して色成分の抽出を行う[2]。そしてその画素における明度の計算、つまり RGB 値の平均値の算出を行い、濃淡画像を作成する。それを一つの画像の全画素に対して行い、ヒストグラムを作る。作成したヒストグラムに大津法を適用し、閾値を算出する。その閾値を使って、濃淡画像に対して２値化を行って２値画像を出力する。

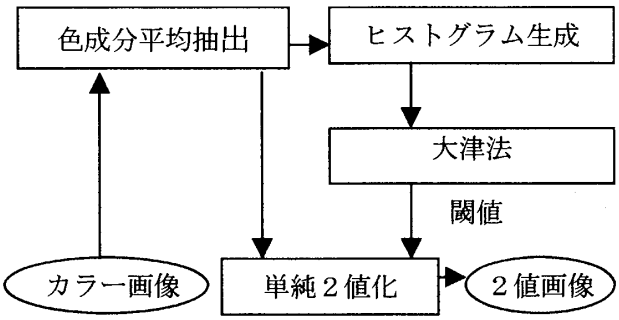


図１ 文字パターン抽出のブロック図

4. 実験内容

実験環境として、CPU 2.40GHz,メモリ 1.25GB, Windous XP Home Edition,スキャナ,EPSON GT－7000（カラー，解像度 200dpi）を用いた。

4.1 実験方法

カラー画像ファイルは図２に示すように読み込まれ、三原色成分を平均化した濃淡画像を生成する。次いで、2 値化において閾値の決定には大津法を組み込んで自動的に２値化を行う。

† 明星大学理工学部, Faculty of Physical Sciences and Engineering, Meisei University

なお、画像ファイルの形式はｐｐｍファイルである。このプログラムはすべてＣ言語で構成されている。プログラムの規模は８６５ステップであった。

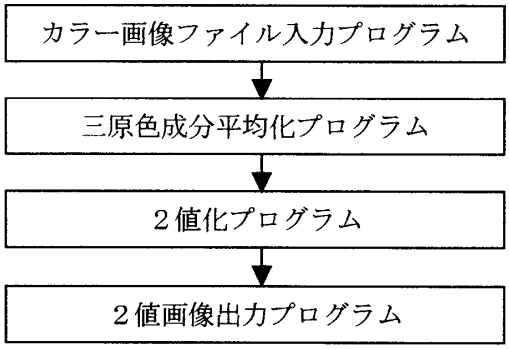


図２　実験手順

4.2　実験サンプル

6 種類のカラー帳票をサンプルとして実験を行った。帳票はスキャナで取り込んでおり、各色成分は 8 ビットである。様々な色の帳票をサンプルとして選び、実験では偏りが出ないようにした。模様のある背景の中にＩＤ部を現す文字列が存在する。帳票画像はおよそ、1300×600 画素である。

4.3　実験結果

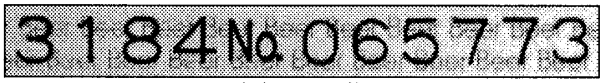
薄めの色が多い場合は特に問題なく文字列の取り出しに成功したが、濃い色を使った場合や背景に文字を印刷してある場合は背景にノイズが残った。表１は２値化結果画像の画質評価であり、評価は 5 人の目視で行った。

表１　実験結果画像の画質評価

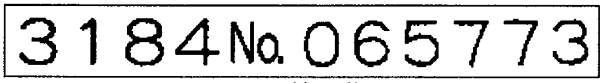
サンプル番号	評価	画像の特徴
1	○	赤メインの色調
2	○	茶系統がメイン
3	○	薄い青
4	×	濃い模様あり（赤系統）
5	△	濃い模様あり（青系統）
6	×	背景に黒ドット散在

※○：良好　△：ノイズ少　×：ノイズ多

図３及び４は、原画像と２値化結果画像である。図３で示すサンプル１は良好な結果が得られているが、図４で示すサンプル４の 2 値化結果では原画像で色の濃い部分が黒い点の塊が散在してノイズとして残っている。これは文字の読み取りの障害になる。図５はサンプル１画像のヒストグラムである。

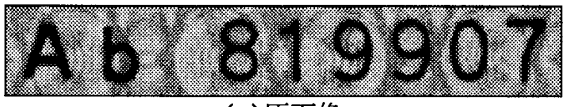


(a) 原画像

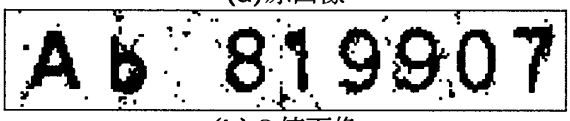


(b) ２値画像

図３　２値化結果（サンプル１）



(a)原画像



(b) ２値画像

図４　２値化結果（サンプル４）

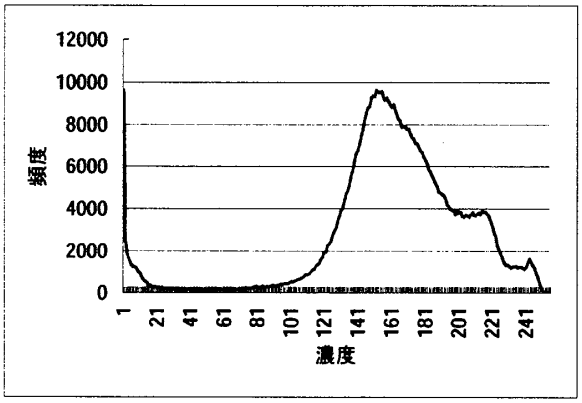


図５　ヒストグラムの例（サンプル１）

５．　まとめ

本方法は画像の２値化として大津法を採用した。三原色成分の平均値を対象として変換を行った結果、薄い色の背景の場合は文字列を取り出すことができ、良好な結果となった。しかし、まだ複雑な背景模様の中からの取り出した文字列にはノイズが残るサンプル画像が存在する。今後、大量のサンプルでの評価実験が必要である。

参考文献

[1] 大津展之: "判別および最小２乗規準に基づく自動しきい値選定法",信学論(D), Vol.63-D, No.4, pp.349--356 (1980-04).

[2] David A.Forsyth, Jean Ponce, "Computer Vision A Modern Approach" ,Prentice Hall , pp.105-113 (January,2003)