配送伝票からの重ね書き文字抽出の検討

1K－1

松尾賢一*　上田勝彦*　梅田三千雄**

*奈良工業高等専門学校　情報工学科　**大阪電気通信大学　情報工学部

1. はじめに

宅配会社は，荷物を配送地域への仕分けの際に区分けコードを使用している．区分けコードは，4桁の数字とその中間にハイフンを持ち，配送伝票上に赤色のフェルトペンで重ね書きされている．本研究では，この配送伝票からの重ね書き文字の抽出手法を提案し，その手法の有効性の評価および検討を行なう．

2. 配送伝票に書かれた区分けコードの特徴

本研究では，図1に示すような宅配用の配送伝票領域が切り出されたものと仮定し，CCDカメラよりRGBの各輝度値8bit，横512×縦400画素の画像サイズで入力した配送伝票を対象として，その上に重ね書きされた区分けコードの抽出を目的としている．この区分けコードは，赤色のペンで書かれていることに着目し，配送伝票上の区分けコードを色彩情報を用いて抽出を行なう．

まず，色彩情報を得るために入力画像にHSV変換を施す．HSV変換によって得られた色の種類を表す色相値(Hue)を横軸に，色の鮮やかさを表す彩度(Saturation)を縦軸にとったH-S分布図を作成する．入力画像の画素ごとに色相値Hと彩度値Sを求め，H-S分布図にプロットする．このとき，図1のH-S分布図の色相値0°付近において，区分けコードの文字色である赤色のクラスタ分布が確認できる．したがって，赤色クラスタに対応した入力画像の画素を選びだすことによって，区分けコードの抽出が可能となる．

しかし，区分けコードは配送伝票上に重ね書きされているために，ボールペンで書かれた文字や伝票枠の印刷色との混色，カメラ入力時における伝票表面上の乱反射などが生じ，本来の区分けコードの赤色クラスタが形成されないという問題が生じる．

そこで，20枚の配送伝票を用いた予備実験によりH-S分布図における区分けコードの色クラスタの分布範囲を明らかにする．この実験により，以下のような知見が得られた．

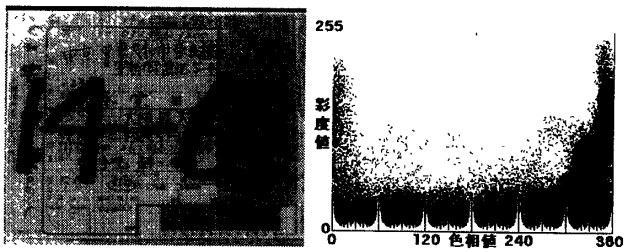

図1　配送伝票とそのH-S分布図

・クラスタの重心は彩度値30以上に存在する．
・クラスタは色相値330°から30°内に分布する．
・クラスタの分散値は均一でない．
・区分けコードのクラスタの出現頻度は高い．

この知見から彩度値と色相値から色クラスタを得ることで，区分けコードの抽出が可能であると考えられる．

3. 区分けコード抽出手法の提案

得られた知見から，3種類の区分けコード抽出手法を提案し，各手法に基づく抽出実験を行なう．ここで提案する抽出手法は，

1. パラメータによる抽出
2. 代表色クラスタを用いた抽出
3. 平均隣接数[1]を用いた抽出

の3種類で，各手法について以下に述べる．

3.1 パラメータによる抽出

この手法は，H-S分布図に対して色相値および彩度値の範囲をパラメータとして与え，パラメータの範囲内に分布するクラスタを調べ，クラスタに対応した入力画像の色領域を区分けコードとして抽出するものである．本手法で与えるパラメータは，知見で得られた値330°$\leq H \leq$ 30°，$s \geq$ 30°を用いる．

3.2 代表色クラスタによる抽出

この手法では，H-S分布図の領域内の各クラスタについて，その出現頻度を調べる．このとき，最も出現頻度が高いクラスタを区分けコードのクラスタとして，そのクラスタに対応した入力画像上の色領域の抽出を行なう．

Extraction of Overlape Writing Characters from Delivery Slips.

Ken-ichi MATSUO*,Katsuhiko UEDA*,Michio UMEDA**

*Nara National College of Technology.

**Osaka Electro-Communication University.

具体的な処理として，H-S 分布図に対して色相値と彩度値をある区間ごとに分割し，粗メッシュ化を施す．メッシュ化することにより，ある程度の色彩変動の吸収と雑音の影響を低減することができる．

このとき，対象となるメッシュ内のクラスタの出現頻度がその周辺のそれより高い場合において，対象メッシュとその 8 近傍メッシュ内に存在するクラスタを入力画像における代表色とする．得られた代表色の中で最も出現頻度の高い代表色を区分けコードの色クラスタとする．

3.3 平均隣接数を用いた抽出

前述した両手法では，適切なパラメータの設定が必要であったり，画像データによっては，抽出に最適な画像が得られないことが考えられる．この理由として，光源の変化や区分けコードの書きムラなどがあげられる．したがって，光源の影響を受けにくい色相値に対し，光源の照度が変化すると彩度値方向にクラスタが移動する．また，区分けコードの書きムラにより文字線色の均一性が失われ，クラスタが分散し，代表色を決定しても周辺のクラスタの範囲を決定するのが困難となる．

そこで，H-S 分布図の知見で得られた色相値の範囲 $330^\circ \leq H \leq 30^\circ$ において，彩度値を変化させた範囲に分布するクラスタに対応した入力画像上の領域を得る．この領域の単純さを平均隣接数で評価を行ない，この平均隣接数が最も高い値である時に得られた領域を区分けコードとして抽出する．

4. 区分けコード抽出手順

各提案手法より，得られたクラスタから入力原画像に対応した画素を画素値の'1'，それ以外を'0' とした 2 値画像を作成し，区分けコードと背景領域とに分離する．画素値'1' の領域に対してラベルづけを施し，それと同時にラベルの面積を求める．ラベル領域が区分けコードと同数のとき，抽出処理を行なうが，ラベル数が区分けコードより多いとき，区分けコードが分離していることが考えられる．

これに対して，膨張・収縮処理を施すことにより，分離切断された連結成分を一つの連結成分とすることができ，それぞれの連結成分の面積が閾値以下の領域を排除することで文字領域以外の不要な領域を削除する．

また，区分けコードよりラベル数が少ないときは区分けコードの接触が考えられる．そこで，接触文字である連結成分に対して，X 軸に連結成分数のヒストグラムをとり，ヒストグラムの谷から各文字ごとの切り出しを行なう．この処理で得られたラベル領域の外接矩形を得ることで，区分けコードが抽出される．

5. 実験結果

上述した手法により，区分けコード抽出実験を行なった．51 枚の配送伝票に重ね書きされた 204 文字に対して，平均隣接数を用いた抽出手法から 204 文字全て抽出でき，抽出率は 100%であった．各手法における抽出結果を表 1 に，抽出結果を図 2 に示す．

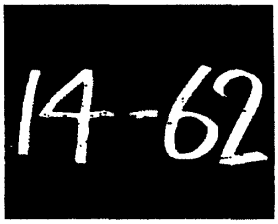

図 2 区分けコード抽出結果例（平均隣接数）

表 1 各手法における抽出結果 (文字総数 204 文字)

抽出手法	抽出文字数	文字抽出率
パラメータ	140 文字	68.6%
代表クラスタ	203 文字	99.5%
平均隣接数	204 文字	100.0%

パラメータによる抽出では，区分けコードのクラスタと色相が類似した伝票の枠線のクラスタが，パラメータ範囲内に共に存在することがあり，区分けコードのみを良好に抽出できる画像が得られず，区分けコード以外の領域が雑音として多く発生した．

代表色による抽出では，良好な抽出結果であるが，代表色とその周辺色の範囲の設定によっては，区分けコードのかすれ，欠落などが見られた．また，線幅の細い文字やシェーディングによって，抽出文字の品質が良好でない文字が若干見られた．

平均隣接数による抽出では，区分けコード以外の領域の発生を極力抑える事ができ，また，区分けコードにおける色彩の変化に対しても，良好な抽出結果を得る事が可能であった．

6. おわりに

配送伝票に重ね書きされた文字の抽出に対し，区分けコードの特徴について調べた．これによって，得られた知見から三つの区分けコードの色クラスタの抽出手法を提案し，各手法に対して区分けコード抽出実験を行ない，平均隣接数を用いた抽出手法において，100%の抽出率を得ると同時に提案手法の有効性を示した．

今後は，抽出された区分けコードの認識についても検討して行きたい．

参考文献

[1] 笹川, 黒田, 池端"平均隣接数に着目したしきい値決定法", 信学論,Vol.J73-D2,No.3,pp.360-366(1990.3)