

3C-8

表の領域情報に基づく帳票識別法の検討

徳升厚美 岩城 修

NTTデータ通信株式会社

1. はじめに

オフィス等で扱われる帳票の自動入力を指向して、帳票の認識技術の確立が望まれている。帳票は、その種類によって読み取るべき文字の位置や認識結果の処理法が異なるため、認識処理に先だてこれを識別する必要がある。本報告では、帳票中に一般的に含まれる表の領域情報を用いて帳票の種類を識別する手法を提案し、また、提案手法に基づいて識別実験を行った結果について述べる。

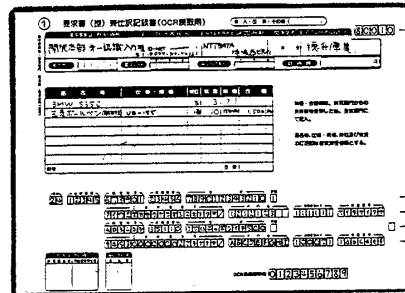
2. 着目する帳票の性質

一般に、帳票はプレ印刷された文字および罫線と、入力データである文字から成る。罫線は文字の記入位置を示す表を構成しており、帳票の種類によってこの表領域の数および位置、大きさが異なることから、これらに関する特徴を抽出し入力帳票の種類を識別する。

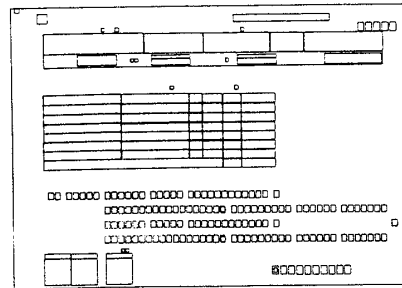
3. 帳票の識別手法

3.1 表領域の抽出法

文字等の記入位置を示す表領域は、罫線に囲まれた矩形領域である。帳票を2値画像として入力した場合、帳票中の個々の表領域は白画素が連続した領域で表されることに着目し、白画素の連結領域を抽出して、これに外接する矩形枠(以下、白連結矩形と呼ぶ)を求める。ここで、白連結矩形には表領域の他に文字中の閉領域から抽出されたものも存在するため、適当なしきい値を用いてこれらの小さい白連結矩形を取り除く。図1に、入力帳票の例と抽出された白連結矩形を示す。



(a)入力帳票



(b)白連結矩形

図1 入力帳票とその白連結矩形の例

3.2 特徴ベクトルの抽出法

先に求めた白連結矩形について、その数および位置、大きさを表す特徴を求め、これを特徴ベクトルの形で表現する。

(1) 位置の特徴

白連結矩形が帳票の上下や左右のどこにあるかを表すため、入力帳票を縦および横にそれぞれ短冊状にM等分し、それぞれの領域毎に含まれる白連結矩形の特徴を求める。

(2) 大きさの特徴

白連結矩形の大きさを表すため、各白連結矩形の縦および横方向の長さをそれぞれ入力帳票の大きさに対する割合で表し、これをN段階に量子化して求める

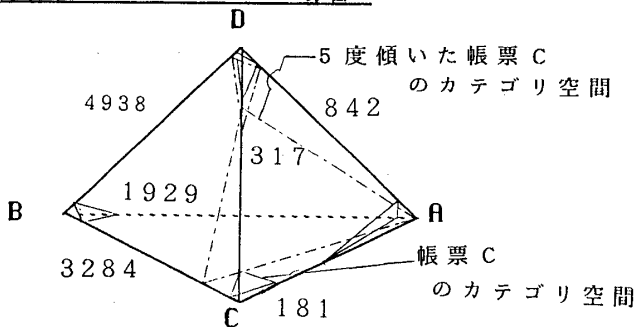
(3) 特徴ベクトル

(1)で求めた各短冊領域について、(2)で求めた各大きさの白連結矩形が幾つあるかを求める。但し、縦および横方向について各分割領域を白連結矩形が完全に占めている場合を8とし、一部のみ存在する場合はこれに領域幅に対する割合を掛けて加算する。ここで、特徴ベクトルの次元数Dは、

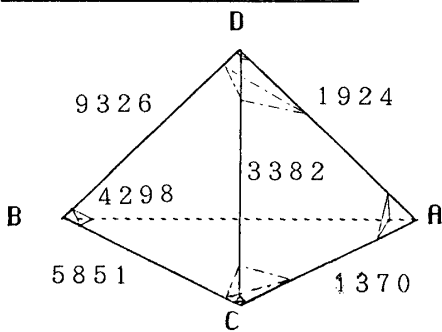
$$D = (M \times 2) \times N^2$$

となる。

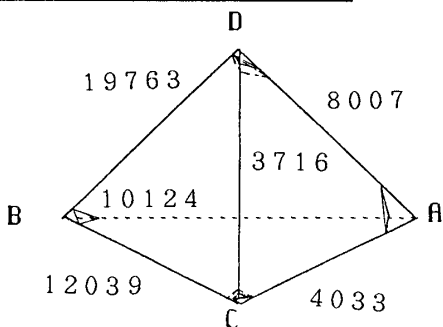
(A) M = N = 1 の場合



(B) M = N = 2 の場合



(C) M = N = 4 の場合



(図中の数値は、帳票A, B, C, Dの学習サンプルの各平均ベクトル間の距離)

図2 各カテゴリ間の距離

4. 識別実験の結果

実験は、4種類の帳票A, B, C, Dをそれぞれ10枚づつ用い、8画素/mmの解像度で入力して行った。具体的には、各種類について7枚を学習サンプルとして平均ベクトルを求め、残り3枚の未知サンプルを含む各サンプルについて、平均ベクトルからのユークリッド距離が最も近いものを正解とする識別実験を行った。

M, Nを変化させ、それぞれの識別能力を求めた結果を図2に示す。図2の(a)はM = N = 1の場合、(b)は同じく2の場合、(c)は4の場合で、4つのカテゴリ間の平均ベクトルの距離と、各カテゴリ内で平均ベクトルから最も遠いサンプルの距離を求め、各カテゴリ空間を実線で示している。この結果、M = N = 1, 2, 4の全てにおいて学習サンプル、未知サンプルとも100%識別できることが分かった。

次に、帳票CおよびDの未知サンプルをそれぞれ1枚づつ5度傾けて入力し、識別実験を行った。その結果、同じく図2の一点鎖線で各カテゴリ空間を示したように、M = N = 1の(a)の場合帳票Cを帳票Aと誤ったが、M = N = 2以上の(b)および(c)の場合は正しく識別できることが分かった。

5. むすび

本報告では、帳票中の表領域を表す白連結矩形を抽出し、これらの数、位置、大きさを表す特徴ベクトルを求めて学習することにより、未知の帳票を識別する手法を提案した。

今後は、罫線のかすれ等により正しく表領域が抽出できない場合や、また本報告の中で一部試みた入力帳票が傾いた場合について検討するとともに、表領域の位置を表す分割数や大きさの量子化数等の最適化を図る。

最後に、本研究を進めるに当たり御指導頂いた開発本部荒川弘照第二技術部長、木田博巳主任技師に感謝する。