コンピュータビジョン　58－16
(1989. 1. 27)

フィールド情報に基づく帳票識別の一検討

A Study of Document Form Identification based on Table Structure

徳升厚美　　岩城　修
Atsumi TOKUMASU　Osamu IWAKI
ＮＴＴデータ通信株式会社
NTT DATA COMMUNICATIONS SYSTEMS CORPORATION

あらまし　本報告では，入力帳票の書式を自動的に識別する手法として，帳票中に一般的に含まれる表領域の数および位置，大きさに関する特徴の違いに着目する手法を新たに提案する．具体的には，表領域を白画素の連続した矩形領域として抽出し，この白連結矩形の数，位置，大きさの特徴を表す特徴ベクトルを求めて学習し，未知の入力帳票を識別する．また，提案手法に基づく実験を行った結果，４種類の帳票に対して100%の識別率が得られ，さらに，傾斜した入力帳票に対しても有効であることが分かった．

Abstract　*This paper proposes a method to identify the document format through the table structure included in the document. Since a table comprises several fields separated by line segments, the proposed method uses connected components of white pixels to extract individual fields. Inputed documents are classified by distance calculation on the feature vector space which represents the number, the position and the size of connected components of white pixels. The experimental result for 4 categories of documents shows that the proposed method is effective even for skewed documents.*

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，認識処理に先だってこれを識別する必要がある．これまで，帳票の固定位置に識別のためのＩＤをプレ印刷し，これを認識して識別するなどの方法が用いられてきたが，既存の多種多様な帳票を処理対象とするには，これらの手法が適用困難な場合がある．そこで，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を識別する手法を提案し，さらに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．複数種の帳票の中から入力帳票の種類を識別するには，固定データである前者の文字または罫線に着目することが有効である．特に，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．そこで，入力帳票の表領域のこれら特徴を表す特徴ベクトルを求め，

学習によって得られた平均ベクトルからの距離によりその種類を識別する手法を以下に述べる．

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．そこで，2値画像として入力した帳票中の個々の表領域が白画素の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るものも存在するため，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，処理対象とする帳票の例として4種類の帳票A，B，C，Dの入力画像と，帳票Bより抽出した白連結矩形を示す．

3．2　特徴ベクトルの抽出法

入力帳票毎に，先に求めた白連結矩形について，その数および位置，大きさに関する特徴を表す特徴ベクトルを求める．

（1）位置に関する特徴

白連結矩形の帳票中の位置に関する特徴を表すため，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白連結矩形の数を求める．ここで，縦方向にM等分したi番目（$i=1,\cdots,M$）の短冊領域をX_i，横方向にM等分したj番目（$j=1,\cdots,M$）の短冊領域をY_jとラベル付けする．

（2）大きさに関する特徴

白連結矩形の大きさに関する特徴を表すため，各白連結矩形の横および縦の長さをそれぞれ入力帳票の横および縦の長さで正規化し，これをN段階に量子化して表す．すなわち，各白連結矩形の横方向の長さをH_k（$k=1,\cdots,N$），縦方向の長さをV_l（$l=1,\cdots,N$）とラベル付けする．

（3）数に関する特徴および特徴ベクトル

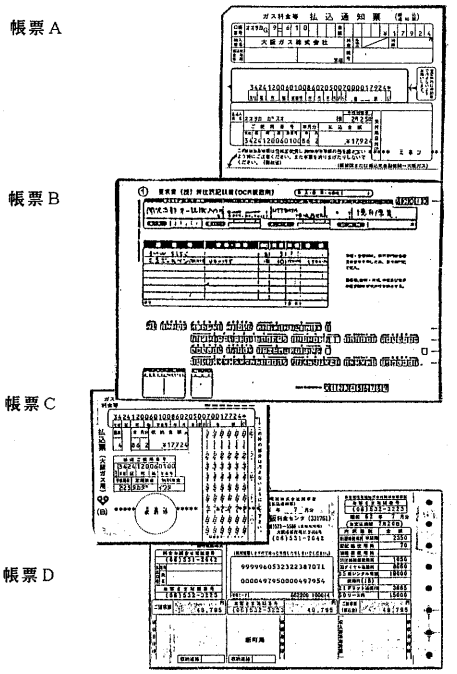
（1）でラベル付けした各短冊領域について，（2）でラベル付けした大きさの白連結矩形の数をそれぞれ求める．ここで，特徴ベクトルの次元数Dを

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

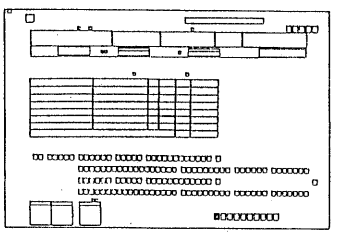
とし，特徴ベクトルの各要素は，

［短冊領域X_iまたはY_jに含まれる横の長さH_k，横の長さV_lの白連結矩形の個数］

とする．



(a)入力帳票



(b)帳票Bの白連結矩形

図1　入力帳票とその白連結矩形の例

4．識別実験

4．1　実験の方法

実験は，図1に示した4種類の帳票A，B，C，Dをそれぞれ10枚づつ用い，8画素／㎜の解像度で入力して行った．具体的には，各カテゴリについて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．以下に，実験で用いたパラメータについて述べる．

（1）分割数および量子化数

帳票の縦および横の分割数M，白連結矩形の縦および横の長さの量子化数Nを変化させ，それぞれの識別能力を比較する．

（2）個数の重み付け

特徴ベクトルの各要素で白連結矩形の個数を求める際，次の3通りの方法で重み付けを行って識別能力を比較する．

① 各白連結矩形の中心を含む領域に8を加算する，

② 各白連結矩形を一部でも含む全領域に8を加算する，

③ ②の方法で，領域の大きさに対し白連結矩形が占める割合に応じて1～8の重み付けを行って加算する．

4．2　実験の結果

（1）識別能力の評価実験

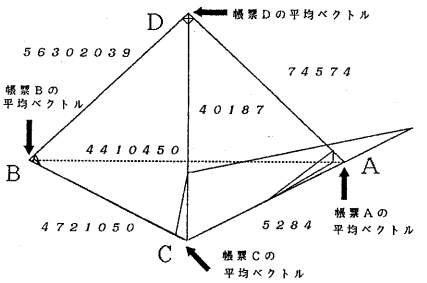
重み付け方法として①の方法を用いた場合の結果を図2に示す．図2で，（a）はM＝N＝1の場合，（b）は同じく2の場合，（c）は4の場合で，4種類のカテゴリ間の平均ベクトルの距離と，各カテゴリ内で平均ベクトルから最も遠いサンプルまでの距離を求め，各カテゴリ空間が占める範囲を実線で示した．この結果，M＝N＝1および2では帳票Aと帳票Cのカテゴリ空間が重なり，識別誤りが生じることが分かった．

また，重み付け方法として②の方法を用いた場合の結果を同じく図3に示す．この結果，M＝N＝1の場合に識別誤りが生じることが分かった．

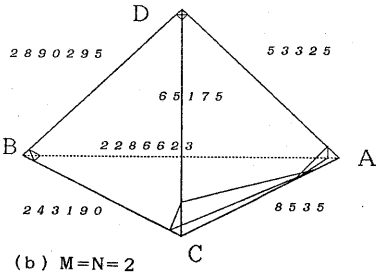
さらに，重み付け方法として③の方法を用いた場合の結果を図4に示す．この結果，M＝N＝1，2，4の全てにおいて学習サンプル，未知サンプルとも識別率が100％となることが分かった．

（2）傾いた帳票に対する識別能力

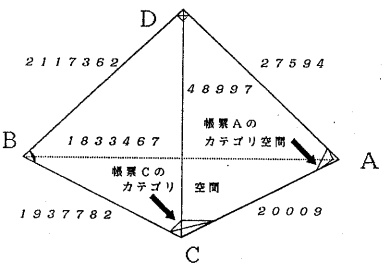
次に，各カテゴリの未知サンプル帳票をそれぞれ1枚づつ5度傾けて入力し，①②③の重み付け方法のうち最も識別能力の高かった③の方法で識別実験を行った．その結果，図4の点線で各カテゴリ空間が占める範囲を示したように，（a）のM＝N＝1の場合識別誤りが生じるが，M＝N＝2，4の場合は正しく識別できることが分かった．



(a) M=N=1



(b) M=N=2



(c) M=N=4

図2　①の方法を用いた結果

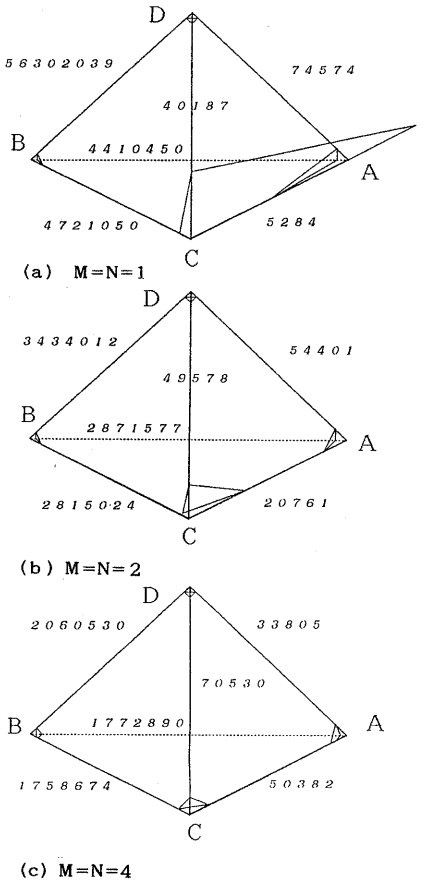


図3 ②の方法を用いた結果

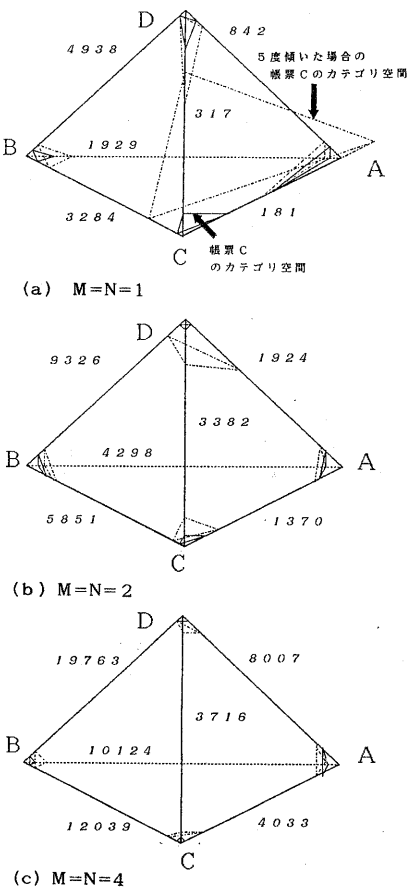


図4 ③の方法を用いた結果

4.3 結果の考察

各白連結矩形を含む全領域で白連結矩形の数をカウントし，さらに白連結矩形が領域を占める割合に応じて重み付けを行うことにより，大きな白連結矩形の寄与度が大きくなり，識別能力が向上することが分かった．

また，分割数Mおよび量子化数Nを小さくすると（例えばM＝N＝1），特徴ベクトルの次元数は小さくなるが，白連結矩形の位置や大きさを詳細に表現できなくなることから，帳票が傾いて入力された場合や，文字中の閉領域が誤って表領域として抽出された場合，および表領域がノイズ等によって分割された場合，識別能力が低下することが分かった．

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，未知の入力帳票を識別する手法を提案した．また，提案手法に基づく識別実験の結果，本手法の有効性を示した．

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帳票や，また本報告の中で一部試みた傾いた帳票についての識別能力を詳細に検討するとともに，帳票の種類，枚数を増やして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り御指導頂いた開発本部荒川弘熙第二技術部長，木田博巳主任技師に感謝する．