H-012

仮説検証型アプローチを用いた定義レス帳票認識技術
Template-free Form Recognition using Hypothesis Testing Approach

平山 淳一　新庄 広　高橋 寿一　永崎 健
Junichi Hirayama　Hiroshi Shinjo　Toshikazu Takahashi　Takeshi Nagasaki

1．背景

業務記録や取引記録を行うための紙帳票は，広く流通しており，多種多様かつ膨大な量の帳票を効率よく処理するための帳票認識技術が求められている．

既存の帳票認識技術は，読取文字の位置や属性を事前にシステムに登録する帳票定義により，帳票画像内の目的の文字を読み取るものが多い．しかし，図１に示すような，同じ種類の帳票でも，番号や日付といった同一項目の読取位置が異なる非定型帳票に関しては，帳票の種類が膨大な数ある場合，全ての帳票種に対し帳票定義を作成することは非現実的である．そのため，帳票定義を作成せずに帳票認識を行う「定義レス帳票認識技術」のニーズが高まっている（例えば関連研究[2]）．本稿では，この定義レス帳票認識技術について報告する．

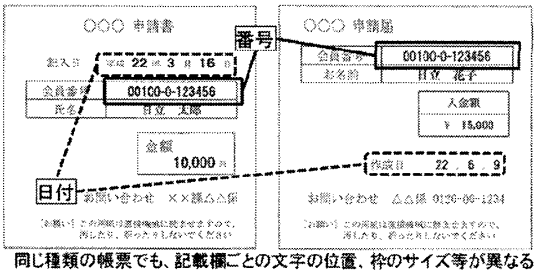

図 1．非定型帳票の例

2．帳票認識技術の概要と課題

2.1. 定義レス 帳票認識技術の従来方式

定義レス帳票認識は，読取項目の位置の事前定義なしに，帳票画像から読取項目を自動的に抽出する技術である．本稿では，読取項目を項目名(Label)と項目値(Value)の２つの概念に分け定義する．項目名とは「番号」や「記入日」といった読取項目の属性に係る文字列，項目値とは「00100-0-123456」や「平成 22 年 3 月 16 日」といった読取項目の値そのものに係る文字列である．

定義レス帳票認識の先行研究として，我々は文献[1]に示す技術を提案している．文献[1]の技術では，まず帳票画像から項目名を抽出し，抽出した項目名の属する枠の隣接枠から対応する項目値を抽出することで，定義レス帳票認識を実現している．項目名の抽出は，項目名辞書（図２）との単語照合によって行う．これにより，従来，帳票種ごとに作成していた帳票定義に代わり，複数の帳票種に共通な項目名辞書を作成するのみで帳票認識が可能になり，大幅なコスト削減が可能となる．

項目名	属性番号
会員番号	001
登録番号	001
記入日	002
作成日	002
日付	002
登録費用	003
使用料	003
作成者	004
お名前	004

図 2．項目名辞書

† （株）日立製作所 中央研究所，Hitachi Ltd. Central Research Laboratory

2.2. 従来方式における課題

文献[1]の技術は，読取項目の並びが規則的かつ配置が密な「整列表形式帳票」，文字の劣化が少ない「高品質文字帳票」を主な認識対象として考案された技術である．一方，下記の条件下では，読取項目の抽出誤りが頻発するといった課題がある．

Ⅰ．読取項目の並びが不規則かつ配置が疎な帳票（非整列枠形式）

ある項目名に対して取り得る項目値の配置が複数あり，一意に項目値を特定できない．また，項目名-項目値の配置関係が隣接枠，同枠内，枠外近傍文字列と複数種類あり曖昧性が高い．例えば，図３(左)の場合，項目名「会員番号」に対し，項目値候補が隣接枠の「22.11.4」と同枠内の「00000-1-222222」の２通り存在し，曖昧性が生じる．

Ⅱ．文字のかすれ・つぶれが多い（低品質文字）

文字認識誤りにより，項目名の抽出漏れが発生し(項目名辞書との単語照合誤りが発生)，その結果，項目名-項目値関係の抽出漏れに直結する．図３(右)の場合，項目名「会員番号」が抽出漏れとなると，対応する項目値の探索を行わないため，項目名-項目値関係の抽出漏れにつながる．

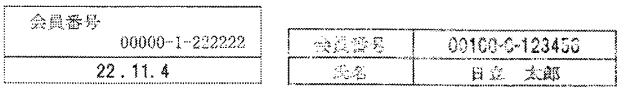

図 3．非整列枠形式(左)，低品質文字帳票(右)の例

上記の課題は，文献[1]の方式が決定論型の項目名-項目値関係探索アプローチ，すなわち抽出が確定した項目名に対し，その周辺から項目値を探索する方式であることに起因して発生する．本稿では，帳票画像内の全ての文字列ペアに対し，それらが項目名-項目値関係にあるスコアを算出し，網羅的にペアの尤もらしさを検証する仮説検証型アプローチを提案する．

3. 仮説検証型定義レス帳票認識技術

3.1. 提案方式の処理フロー

まず，帳票内の全ての文字列ペア (S_i, S_j) について，これらが項目名-項目値関係として妥当な配置関係であるかを表す配置スコア $Salign(S_i, S_j)$ を計算する．配置スコアは，文字列矩形の配置関係（座標関係）のみから計算する．

次に，全ての文字列 S_i, S_j について，当該文字列が辞書に登録された単語であるかを表す項目名スコア $Slabel(S_i, W_n)$ および項目値スコア $Svalue(S_j, D_m)$ を計算する．項目名スコアは項目名辞書の単語 W_n と，項目値スコアは表記辞書 D_m とそれぞれ照合を行うことで計算する．

計算された配置スコアおよび項目名スコア，項目値スコアの統合により項目名-項目値関係を決定する．仮説の検証には，スコアの統合を行う評価関数を用いることとし，評

価関数の値が属性ごとに最大となる文字列ペア (S_i,S_j) を項目名-項目値関係と決定する．評価関数には，

$$(LS+VS)\times AS$$

を用いた．LS は項目名スコア，VS は項目値スコア，AS は配置スコアをそれぞれ表す

3.2. 配置スコアの計算

配置スコアは $Salign(S_i,S_j)$，2つの文字列矩形の配置関係，文字列が属する枠の隣接関係，枠内での文字列矩形の位置などから決定する．文字列ペアの配置関係は，隣接枠，同枠内，枠外近傍の3つに場合分けできる．2つの文字列の配置の項目名-項目値関係としての妥当性の定量化は，妥当な文字列矩形配置からのずれをペナルティとして付加することで実現できる．本方式では，2つの文字列矩形配置の非妥当性をペナルティとして定量化した．

$$Salign(S_i,S_j)=\max\{1-\sum_{k=1}^{K}\alpha_k g_k(S_i,S_j),0\}$$

$g_k(S_i,S_j)$ は配置の非整然性を表したペナルティ関数，α_k は係数である．帳票サンプル内の項目名-項目値関係の配置パターンを分析した結果を元に，ペナルティのルールを表１の6種類に分類した（K=6 ）．

例えば，2つの文字列が隣接枠関係にある場合（表１の上から2つの項目），”項目名文字列は枠の中心にある”，”項目名枠の高さよりも項目値枠の高さが大きい”といったルールから外れた場合に付加される．

表 1．配置スコア計算に用いるペナルティパターン

配置関係パターン	ペナルティパターン	項目名と項目値の配置例	ペナルティ計算式
隣接枠	項目名文字列が枠の中心座標から離れている場合	d1 d2	$g1(i,j)=d1\ d2$
	項目値枠の高さより項目名枠の高さが大きい場合	h1 h2	$g2(i,j)=h1\ h2$ (if $h1>h2$)
同枠内	項目名文字列より項目値文字列が左側(上側)にある場合	d2 d1	$g3(i,j)=d1\ d2$
枠外近傍文字列	項目名文字列高さと項目値文字列高さが異なる場合	h1 h2	$g4(i,j)=h1\ h2$ (if $h1>h2$) $g4(i,j)=h2\ h1$ (if $h1<h2$)
	項目名文字列と項目値文字列が離れている場合	w1 d1	$g5(i,j)=d1\ w1$ (if $d1>2\times w1$)
	項目名文字列より項目値文字列が左側(上側)にある場合	d2 d1	$g6(i,j)=d1\ d2$

3.3. 項目名スコアの計算

項目名スコア $Slabel(S_i,W_n)$ は，帳票画像内の文字列 S_i が項目名辞書内の項目名単語 W_n である確率を表した尺度であり，候補文字ネットワークの文字列パス探索結果から算出される．

$$Slabel(S_i,W_n)=\frac{1}{C}\sum_{c=1}^{C}Clabel(S_i,W_n)_c$$

但し，$Clabel(S_i,W_n)_c$ は項目名辞書内の n 番目の単語 W_n の c 番目の文字に対する個別文字の識別スコア，S_i は帳票内の文字列，C は項目名単語長である．

提案方式では文字認識誤りへの頑健性を向上させるため，誤・不読を許容した単語照合を行う．図４のパス２では辞書に登録された単語「作成者」を候補文字ネットワークから探索するが，このときに「者」が該当ノードの文字識別候補になくとも強制的に照合できた（「？」のノードを通過した）として，文字列パスを決定する（$Clabel(S_i,W_n)_c=0$ とする）．強制照合文字数は，項目名単語長の半分以下とした．

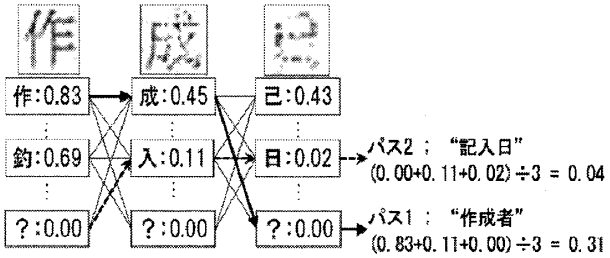

図 4．項目名スコアの計算例

3.4. 項目値スコアの計算

項目値スコア $Svalue(S_j,D_m)$ は，帳票画像内の文字列 S_j が，属性ごとに用意した表記辞書 D_m に記載された表記文法（正規表現）に一致する確率を表した尺度である．項目名スコアと同様に，候補文字ネットワークからの文字列パス探索時の，各文字の識別スコアの平均と定義した．項目名スコアと同様に誤・不読許容型の単語照合を行う．強制照合文字数は，文字列パス長の1/4以下とした．

4．評価実験およびまとめ

帳票サンプル200 枚(200 種）を対象に，帳票読取実験を行った．帳票画像は200dpi の2 値BMP，帳票内平均文字列数は300 であった．１サンプル内の平均読取項目数は5である．項目名-項目値のペアごとの読取率を表２に示す．決定論型の従来方式と比較し，再現率が２倍に向上した．

表 2．決定論型と仮説検証型の認識精度の比較

	従来方式（決定論型）	提案方式（仮説検証型）
再現率(Recall) [%]	36.7	74.4
適合率(Precision) [%]	65.2	83.5

表３に従来の決定論型アプローチにて，項目名-項目値関係抽出誤りが発生していた帳票サンプルでの読取結果を示す．非整列枠形式および低品質文字帳票において，正しい抽出結果が得られた．

表 3．新たに正読となった書式例

	非整列枠形式	低品質文字
画像例	会員番号 00000-1-222222 / 22．11．4	00100-0-123456 / 太郎
従来方式（決定論型）	抽出失敗（対応づけ誤り）“会員番号” - “22．11．4“	抽出失敗（未抽出）“???” - “???”
提案方式（仮説検証型）	抽出成功 “会員番号” - “00000-1-222222“	抽出成功 “会員番号” - “00100-0-123456“

以上の結果より，仮説検証型の項目名-項目値関係抽出方式が，文字認識誤りおよび配置の曖昧性にロバストな帳票認識技術となることを確認できた．今後は，更なる再現率の向上，処理時間の削減が課題である．

文　　献

[1]新庄広，関峰伸，丸川勝美，永崎健，中島和樹， 特開2008-204226“帳票認識装置およびそのプログラム”
[2]宇田明弘，”表形式既存帳票認識システム” FIT2002