領域指定に対してロバストな帳票内文字列読取方式

関　峰伸† 池田　尚司† 酒匂　裕† 今泉　敦博‡
†(株)日立製作所　中央研究所 ‡(株)日立製作所　情報機器事業部

1. はじめに

金融機関や官公庁の窓口では，紙の帳票を用いて税公金や保険料等，様々な振込，納付，申請の業務が行われている．そして，この帳票を用いた確認，承認，修正等の事務処理と記載内容の電子化に多くの時間が費やされている．これら業務の効率化のためには，窓口や ATM 等で帳票を電子イメージ化し，振込先，振込み金額等の記載内容を機械で読取り電子化する帳票イメージ処理技術する必要がある．本報告では帳票イメージ処理における項目読取方式について述べる．

2. 帳票イメージ処理の概要

帳票イメージ処理の目的は，帳票に記載された項目を機械で読取ることである．銀行で取り扱う税公金の振込帳票は数百種類/支店あり，読むべき項目の位置は帳票種毎に異なる．そこで，予め各帳票種のレイアウト情報と読取るべき項目の位置を登録したデータベース(帳票定義と呼ぶ)を用意しておく。そして，入力された帳票がどの帳票なのか知るために帳票種の識別を行う．そこでは，帳票定義のレイアウト情報と入力イメージのレイアウト情報を照合し，帳票種を識別する[4]. 次に帳票定義に登録してある読取るべき項目の領域を切り出す．そして，切り出された領域から項目の読取りを行うことで，金額や納期限等を自動的に電子化する．以降，項目の読取りについて述べる．

3. 課題とアプローチ

3.1. 複雑背景下の項目読取り

帳票定義には予め読取るべき項目の領域(定義領域)が設定されている。しかし，帳票用紙サイズや印刷のばらつき，画像内にある帳票の伸縮や位置検出の誤差により読取り領域がずれてしまう場合がある．そのため，定義領域よりも拡大した領域(拡大領域)から項目を読取る必要がある．しかし，帳票イメージ処理では様々な帳票を取り扱うため，枠線の形式，色，太さ，下地の色，文字の色，フォントは帳票種毎，そして一枚の帳票内でも場所毎に異なる．近傍には他の文字列やプレ印刷文字が存在する．そのため，拡大領域中には読取対象の文字列以外に様々な濃淡・形状を持つパタンが含まれてしまう(図 1)．このような読取りに対して障害の多い背景を複雑背景と呼ぶことにする．複雑背景下での項目読取りにおける課題は，課題１:文字成分の抽出，課題２:文字列の選択の二つに大別できる．

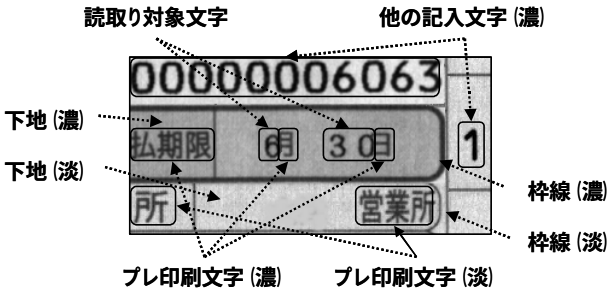


図 1　複雑背景

3.2. 課題１：文字成分の抽出

文字成分の抽出では，濃淡値の違いにより文字成分と他の成分を区別し文字成分のみを黒画素となるように２値化する．図 2 は多値画像(図 1)の濃淡分布である．この全体の濃淡分布において，従来法である k-means 法によるクラスタリングを行い，得られた２値化閾値は閾値Ａと閾値Ｂとなる[1]．そして，その２値化結果は図 3 である．読取り対象文字は，閾値Ａでは黒く塗り潰され，閾値Ｂでは消えているため，読取ることができない。このように読取り領域中に様々な濃淡値の文字，プレ印刷文字，枠線，下地の濃淡色が存在する場合，それらを区別し，文字成分のみを抽出することは難しい．そこで，どのように文字成分を抽出するかが課題となる．我々は，これに対して枠領域分割による文字成分抽出方式を提案する．

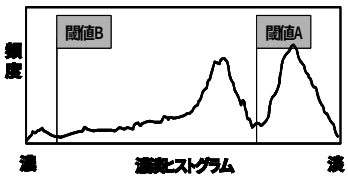


図 2　拡大領域の濃淡分布

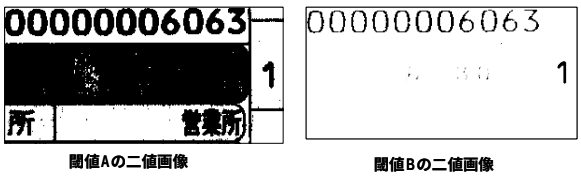


図 3　閾値Ａと閾値Ｂの２値画像

3.3. 課題２：文字列の選択

拡大領域には，読取り対象の文字以外にも様々な文字が混入する．そのため，拡大領域からは複数の文字列が抽出される．そして，どの文字列が読取り対象であるかは帳票種毎，場所毎に異なる．そこで，抽出された複数の文字列の中からどれを選択するかが課題と

A Robust Form-data Recognition Method from different underground
Minenobu Seki†, Hisahi Ikeda†, Hiroshi Sako† and Atsuhiro Imaizumi‡, †Central Research Laboratory, Hitachi, Ltd., ‡Mechatronis Systems Division, Hitachi, Ltd.

なる．我々は，これに対して項目知識利用文字列選択方式を提案する．

4. 枠領域分割による文字成分抽出

「どのように文字成分を抽出するか」について述べる．帳票内の各項目は枠によって区切られており，1つ1つの枠の中にある文字と下地とプレ印刷文字はそれぞれ一様な濃淡分布で記載されることに着目した．そして，読取り領域全体の濃淡分布は複雑であるが，その中の意味のある構造である枠という単位で濃淡分布をみることにより問題を単純化する．まず読取り領域内を枠構造解析[2]する．そして各枠毎に領域分割を行う．すると各枠領域内の画素は記入文字とプレ印刷文字と下地の3種類が存在する場合と，記入文字と下地の2種類の存在する場合の2通りである．そこで，各枠領域の画素3クラスへのクラスタリングと2クラスへのクラスタリングを行う．一般的に，いずれの場合も読取るべき文字列は最も濃いクラスタである．従って，最も濃いクラスタの画素成分を抽出した．クラスタリングには k-means 法を用いた．

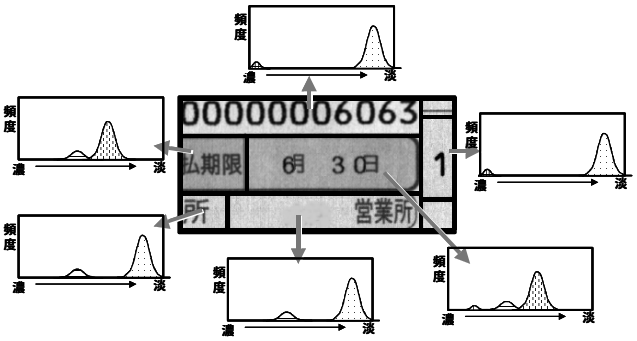


図 4　枠領域分割による文字成分抽出

5. 項目知識利用文字列選択

「どのように文字列を選択するか」について述べる．本方式では，基本的に文字列の読取り結果をもとに文字列の選択を行う．納期限欄を読取る場合には“平成年月日”，金額欄を読取る場合には“￥マークと数字の羅列”というように各読取り欄には固有の表記パタンがあることを利用する(図 5)．まず各文字列の読取りを行う．そして，予め読取り欄毎に用意しておいた表記パターンに合致する読取り結果を得た文字列を選択する．ただし，すべての文字列の読取りを行うと，処理時間が多くなることが問題となる．そこで定義領域(拡大前の領域)が読取り対象文字列を含む枠に沿って指定されることを利用し，次の2つの初期選択を行った．1つ目は形状の情報を利用する．定義領域の矩形に近い枠を選択，すなわち定義領域とは大きく異なる枠を選択対象から除く．2つ目に位置の情報を利用し，読取り領域の中心に近い枠から順に選択し，その中の文字列の読取りを行った．このように表記パタンと形状と位置の情報を利用した文字列の選択を行うことで，高速かつ安定した読取り結果が得られた．

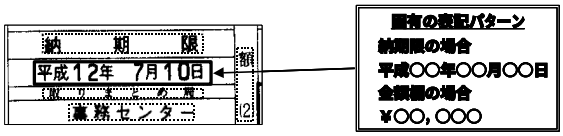
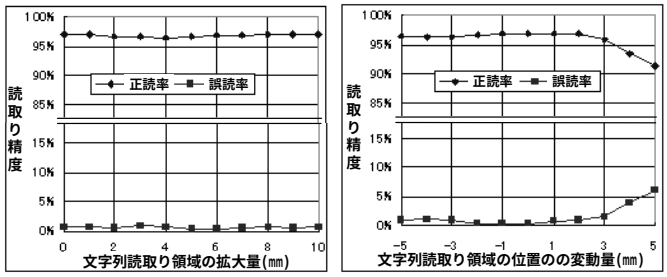


図 5　表記パターンの利用

6. 評価実験および考察

(1)複雑背景に対する頑強性と，(2)文字列読取り領域の位置変動に対する頑強性の評価を行う。評価にはテスト帳票の納期限欄 648 件を用いた．(1)では，読取り領域を徐々に拡大していく際の文字列認識精度の変化を見た．図 6 (a)より，提案方式は 10㎜ の領域拡大でも精度低下を 1pt 以内に保つことができた．(2)では，定義領域から 5㎜ 拡大した読取り領域を水平･垂直方向に 1㎜ ずつずらした場合の認識精度を測定した．図 6 (b)より読取り領域の位置変動が±3㎜ 以内である場合の読取りの精度は 95%以上となった．ただし±3㎜ を超える場合，読取精度は 91%になった．これは納期限の表記と同じ日付文字列が 5㎜ 以内に存在しており，誤った文字列を選択したためである．これに対しては帳票定義を工夫する等のアプローチが考えられる。



(a) 複雑背景と読取り精度　　(b) 位置変動と読取り精度

図 6　複雑背景･位置変動に対する読取り精度

7. おわりに

帳票中の項目読取方式として，枠領域分割による文字成分抽出方式，項目知識利用文字列選択方式を考案した．そして帳票用紙サイズや印刷のばらつき，画像内にある帳票の伸縮や位置の検出誤差による文字列読取り領域の位置変動がある場合でも高い文字列読取り精度を維持することが可能となった．

参考文献.

[1] Oivid Deu Tier, Anbil K.Jain, “ Goal-Directed Evaluation of Binarization Methods”, *IEEE Trans.on.* PAMI, vol.17, no.12, pp.1191-1201, 1995.

[2]H.Shinjo, E.Hadano, K.Marukawa, Y.Shima and H.Sako, “A Recursive Analysis for Form Cell Recognition”, Proc.of ICDAR 2001, pp.694-698, 2001.

[3]H.Sako, M.Seki, N.Furukawa, H.Ikeda, and H.Ogata, “Form Reading based on Form-type Identification and Form-data Recognition”,Proc of ICDAR2003,pp,2003.(to be published)

[4]古川直弘,今泉敦博,藤尾正和,酒匂裕,「星座認識による帳票識別方式」，信学技法, PRMU2001-125, pp85-92, 2001.