4W-6

フォーマット分類機構に基づく帳票管理支援システムの試作

藤江 翔太郎† 白松 俊‡ 大囿 忠親‡ 新谷 虎松‡

名古屋工業大学工学部情報工学科† 名古屋工業大学大学院工学研究科情報工学専攻‡

1. はじめに

見積書，請求書，領収書といった帳票は紙媒体であり，一般家庭における帳票の所有者にとって，膨大な量の帳票の管理は困難である．企業や事務所といった組織では，既存の企業向け帳票管理ソフトウェアが利用できる．個人で帳票を管理する場合では，企業向け帳票管理ソフトウェアは高価なため，導入が困難である．

本稿では，帳票の電子化，閲覧，集計をまとめて帳票管理とする．個人規模の帳票管理は，表計算ソフトを用いた手入力による管理方法がある．コンピュータの扱いに慣れていないユーザにとって表計算ソフトを使いこなすのは困難であり，手入力を行う手間もある．本研究では，タブレット端末を利用した，帳票管理支援システムの開発を行った．タブレット端末は会議などの意思決定の場で利用されるなど，デスクトップパソコンの代替として利用される機会が多く，コンピュータの扱いに慣れていないユーザも手軽に利用できる．

本稿では，帳票管理におけるフォーマット分類機構を提案し，コンピュータの扱いになれていないライトユーザ向けのシステムについて述べる．

2. フォーマット分類機構

帳票管理におけるユーザの操作手順の簡略化を目的としたフォーマット分類機構を提案する．フォーマット分類機構は，電子化を行った帳票の分類や，管理・集計に必要となる文字や数字の取得を自動で行う．

フォーマット分類機構は，「登録」，「分類」，「抽出」の 3 つの機能がある．登録機能は，新規の帳票フォーマットを登録する．分類機能は，電子化した帳票を登録済みのフォーマットに分類する．抽出機能では，電子化した帳票から必要な情報を抽出する．

On a Form Management Support System based on Format Classification Mechanism.

† Shotaro Fujie, Department of Computer Science, Nagoya Institute of Technology.

‡ Shun Shiramatsu, Tadachika Ozono, Toramatsu Shintani, Graduate School of Computer Science and Engineering, Nagoya Institute of Technology.

図 1 にフォーマット分類機構のアルゴリズムを示す．ユーザは，タブレット端末のカメラ機能やイメージスキャナからタブレット端末に取り込んだ帳票の画像データをフォトアルバムから選択する．

```
選択した帳票 : C
if C is 未登録 then
  C の特徴抽出とスコア計算
  C のフォーマットを新規登録
  C 上の情報を抽出
else
  C 上の情報を抽出
end if
```

図 1　フォーマット分類アルゴリズム

取り込んだ帳票が新規の帳票フォーマットかどうかを分類し，登録済みの場合，登録済みのフォーマットに分類する．先行研究[1]として，帳票のセル構造を用いて識別を行う手法がある．本研究では，分類スコアの高い帳票の中からユーザに選択操作を求め，帳票の分類を行う．

2.1 登録

登録とは，帳票フォーマットおよび，管理に必要な情報の登録を指す．帳票の発行元と帳票の種類別に登録を行う．帳票は発行元によってフォーマットは様々であり，同じ発行元の帳票はフォーマットが共通である，発行元が同じでも，領収書や請求書といった帳票の種類によってフォーマットが異なる場合があるため，発行元と種類別に登録を行う．

本研究では，管理に必要な情報を次の 6 つの項目とする．集計に用いる「金額情報」，請求期限などの「期限情報」，発行元を示す「タイトル」，帳票の所有者や対象者を表す「ユーザ」，帳票の「種類」，および，メモなどの「備考情報」である．本機構では，管理に必要な情報のうち，「金額情報」と「期限情報」を自動で取得するために，OCR による文字認識を行う．帳票の発行元が特殊なフォントやロゴで書かれている場合を考慮し，「タイトル」と「備考情報」はユーザによる手入力を必要とする．「ユーザ」および「種類」は一覧から選択をする．「金額

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.

情報」と「期限情報」はすべて値が異なるため，自動抽出することで手間を省くことができる．その他の情報は，帳票フォーマットごとに共通であり，一度だけ登録操作を行えばよい．

帳票上には必要な情報以外に，不要な情報が多くあるため，帳票の画像データ全体を OCR で認識し，必要な情報のみを得るのは困難である．帳票フォーマット登録の際，画像データ上の必要な情報が書かれている領域を指でドラッグして選択することで，必要な情報のみを OCR で認識することを可能にした．ドラッグした領域は座標データとして登録する．

2.2 抽出

登録の際に保存した領域座標データを用いて必要な情報の自動抽出を行う．自動抽出を行う情報は，「金額情報」と「期限情報」である．保存された領域座標データから帳票画像データをトリミングする．トリミングした画像を二値化処理し，OCR で認識する．認識文字列は，表 1 の文字列に限定することで精度を上げる．OCR ソフトウェアには，オープンソースの Tesseract-OCR の iOS ライブラリを用いる．

表 1　認識文字列一覧

金額	0 1 2 3 4 5 6 7 8 9 , ¥
期限	0 1 2 3 4 5 6 7 8 9 / 年 月 日

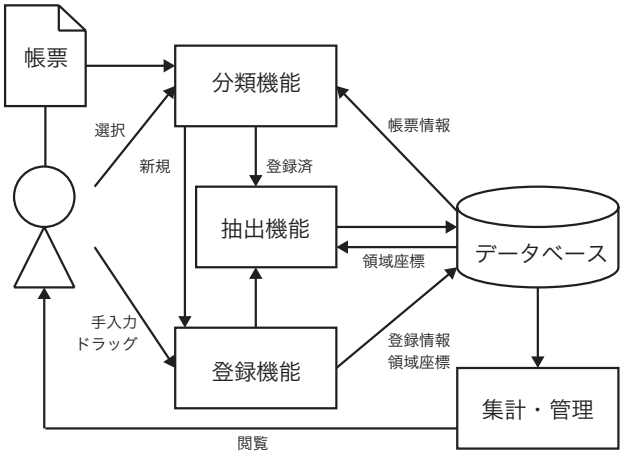

図 2　システム構成図

3. システム概要

図 2 に本システムの構成図を示す．本システムは iPad アプリケーションとして実装した．ユーザは紙媒体である帳票を，iPad のカメラ機能やイメージスキャナを用いて iPad に取り込む．取り込んだ帳票はフォトアルバムに保存される．帳票の画像データをフォトアルバムから選択する．選択された帳票はフォーマット分類機構によって，新規フォーマットか登録済みフォーマットに分類される．新規フォーマットの場合，ユーザは登録操作を行う．登録情報はアプリケーション内のデータベースに保存される．登録済みのフォーマットの場合，登録情報から自動抽出を行う．抽出された情報はデータベースに保存され，集計と管理を行い，ユーザは集計結果や，期限情報を閲覧することができる．図 3 は電子化した帳票の閲覧インタフェースであり，帳票を種類別に閲覧することができる．図中の①は帳票の種類が書かれたヘッダであり，②の部分に対応する帳票のサムネイルが一覧表示される．③は帳票のタイトルを表している．

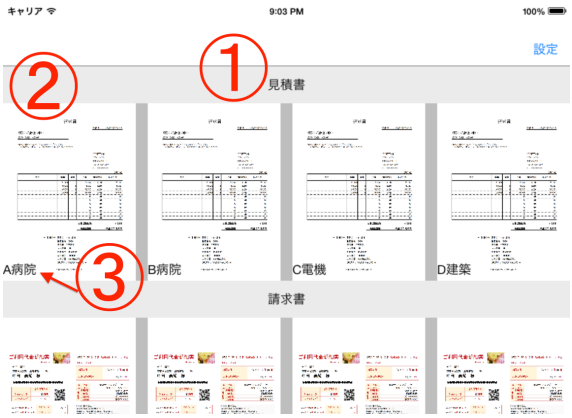

図 3　閲覧インタフェース

4. 評価と考察

本システムの評価実験として，帳票の電子化の処理速度を計測した．10 種類の帳票を新規登録し，同種類の帳票を電子化した．実験には，Apple 社の iPad2(OS:iOS7)を用いた．

帳票フォーマットの新規登録には 1 枚あたり平均 23.5 秒かかった．登録済みフォーマットの電子化には 1 枚あたり平均 2.47[秒]かかった．新規登録には帳票タイトルの入力などユーザによる操作が含まれるため，時間がかかってしまうが，登録済みであればユーザの操作が少なく，短時間で処理が可能である．

5. おわりに

本稿では，膨大な量の紙媒体である帳票の管理を支援する帳票管理支援システムについて述べた．帳票の電子化の際のユーザによる操作数を減らすため，フォーマット分類機構を提案した．フォーマットを登録し，自動抽出を行うことで，ユーザは帳票の電子化のたびに入力を行う必要がなくなる．帳票 1 枚あたりの処理時間の短さから，帳票管理支援の優位性を示した．

参考文献

[1] 浅野他:"セル構造を用いた帳票識別",電子情報通信学会論文誌,Vol. J80-DII, No. 1, pp. 131-138(1997)

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.