読み取り対象を自動的に見付け出す項目サーチOCR

Item-Spotting OCR Technology to Automatically Locate Recognition Targets

鈴木　智久 SUZUKI Tomohisa　　横田　和章 YOKOTA Kazuaki

近年，業務効率化の手段としてRPA（ソフトウェアロボットによる業務自動化）が脚光を浴びており，紙文書を扱う業務でのOCR（光学的文字認識）技術の重要性が増している。しかし，これまでのOCRで帳票から情報を抽出するためには，読み取り対象の位置を様式ごとに指定して作成した“モデル”が必要であり，扱う様式の種類が多い場合や，同じ様式でもレイアウトが一貫していない場合などでは，モデルを用意するのが難しいという問題があった。

そこで東芝デジタルソリューションズ（株）は，指定された位置の文字列を読み取るのではなく，読み取りたい項目を文書中から自動的に見付け出して読み取りを行う“項目サーチOCR”を開発した。この技術を用いることで，多種多様な紙文書を扱うRPAが可能となった。

The emergence of robotic process automation (RPA) in combination with printed and handwritten documents has made optical character recognition (OCR) a key technology for the improvement of productivity in business processes. Since the conventional implementation of OCR requires models specifying the position coordinates and formats of the items to be recognized, time-consuming preparation of these models has been an obstacle to automation. This obstacle is especially pronounced in cases where the number of forms is large or the layouts of the forms are inconsistent and unpredictable.

To enable RPA in such cases, Toshiba Digital Solutions Corporation has developed an item-spotting OCR technology that can automatically find the locations of items targeted for recognition by searching for strings or patterns specified for those items and/or the corresponding labels.

1. まえがき

近年，働き方改革の手段の一つとして，RPA（ソフトウェアロボットによる業務自動化）による業務効率化が求められている。RPAのメインターゲットである事務作業では，電子化により減少したとはいえ，いまだに多くの紙文書を取り扱う必要があり，RPAが効果を発揮するにはOCR（光学的文字認識）による文書の読み取りが重要である。

アプリケーションやRPAソフトウェアなどでOCRを活用するためには，文書中から読み取り対象を抽出する必要がある。従来は，主に文書中の位置情報を用いて読み取り対象の抽出を行っていたが，記載項目ごとにその座標情報を登録する“モデル”の作成に手間が掛かるという問題があった。

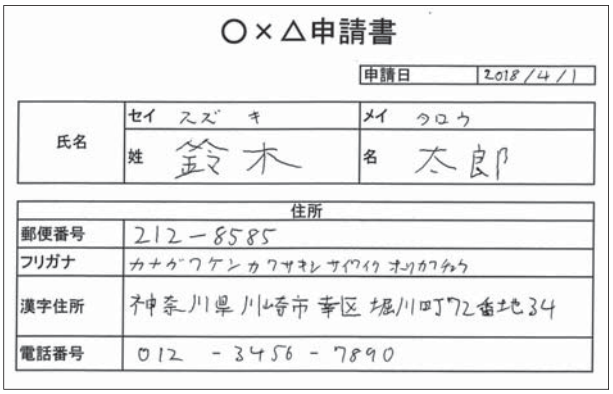
そこで今回，読み取り対象の座標を登録しなくても，読み取り対象を自動的に見付け出す“項目サーチOCR”を開発した。この技術は，読み取り対象の座標の代わりに，読み取り対象の規則や見出しの文字列を登録して，その情報に沿って文書中から読み取るべき文字列の場所を探し出すことができるので，記入項目の座標を登録する必要がない。

ここでは，項目サーチOCRの概要と特長について述べる。

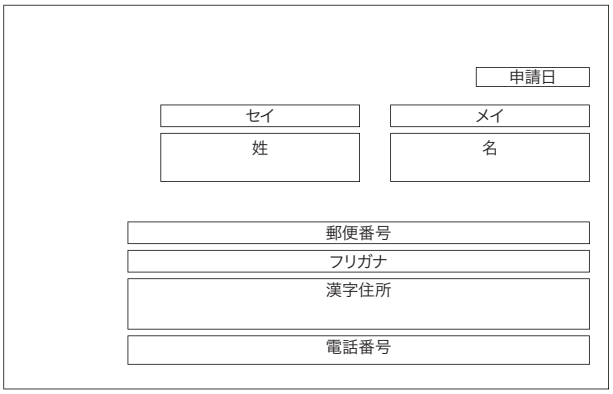
2. 従来型OCRの問題

これまでの帳票OCRでは，**図1**に示すように，モデル登録用のGUI（グラフィカルユーザーインターフェース）で記載項目ごとに位置を指定してモデルを作成する必要があった。読み取り対象の記載項目が異なる場合や，記載項目の位置が異なる場合は，単一のモデルで対応することができない。したがって，記載項目やその位置が異なる場合は，別のモデルを作成し，モデルを切り替えて読み取りを行うのが一般的であった。

モデルの作成には，様式のバリエーション数に応じた手間が掛かる。また，名刺や，領収書，請求書など不特定多数の発行元から受領する書類には，読み取る内容が同じでも記載項目の位置が異なるレイアウトが無数に存在するものがある。このような場合には，モデル作成が困難であるという問題があった（**図2**参照）。



(a) 帳票画像



(b) モデル

図1. モデルへの記載項目の登録
記載項目ごとに位置や識別子などを指定する。
Registration of items to be recognized in model

3. 読み取り対象と見出しのパターン化によるOCR

2章で述べた問題に対応するため，記載項目の座標の登録を前提としない手法が提案されており，大別すると以下の二つの手法が挙げられる。

一つ目の手法は，見出し文字列による記載項目の識別である。記載項目の近くには，固有の見出し文字列が印字されていることが多いため，見出し文字列から記載項目を見付け出す手法が提案されている[(1)]。

二つ目の手法は，記載項目それ自体の字面による識別である。例えば，都道府県名で始まる文字列は住所であると推定でき，このような特定の単語や文字列を検出することで，記載項目を見付け出す方法が提案されている[(2), (3)]。

今回開発した項目サーチOCRでは，前述の二つの手法を取り入れており，見出し文字列と記載項目自体の双方の

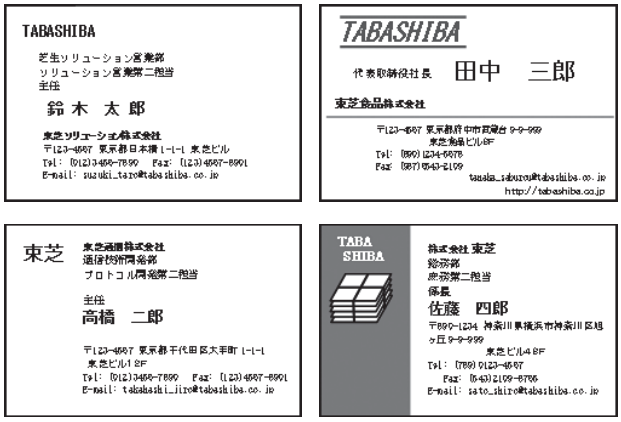


図2. レイアウトのバリエーション
例えば名刺は，発行元によって記載項目のレイアウトが異なるため，レイアウトのバリエーションが無数に存在する。
Example of layout variations

字面を，単一の文字列や，単語リスト，字種指定，正規文法，あるいはそれらの組み合わせで定義し，合致する記載項目を自動的に見付け出すようにした[(注1)]。

文書中の文字列が正規文法やリストに合致しているか否かの判断は，当社が開発した“フレキシブルOCR知識処理[(4)]”により，合致の程度を評価することで行う。フレキシブルOCR知識処理を適用することで，記載項目を自動的に見付け出すことができると同時に，文字認識の誤りを定義に合わせて訂正することも可能である。

項目サーチOCRによる読み取り対象定義の例を，以下に示す。

(1) 見出しで見付け出す記載項目の例　**図3**で示した記載項目の例では，見出し文字列が「お客様番号」であり，記載項目自体が数字だけで印字されていることを規定している。この記載項目は，見出し文字列と合致する文字列を文書中から探すことで見付け出すことができる。

(2) 記載項目自体の字面で見付け出す記載項目の例
図4で示した記載項目の例は，対応する見出しがなく見出しでは探すことができないため，その字面で見付ける。この記載項目は姓及び名のリストに記載された文字列を並べた形となっており，リスト中の単語を手掛かりとして記載項目を見付けることができる。

（注1） 2018年6月現在，東芝デジタルソリューションズ（株）では項目サーチOCRの技術をクラウドサービスとして提供しているが，このサービスでは単語リスト・正規文法による定義は未対応。今後対応の予定。

(a) 見付け出したい記載項目

(見出し)
文字列:
お客様番号

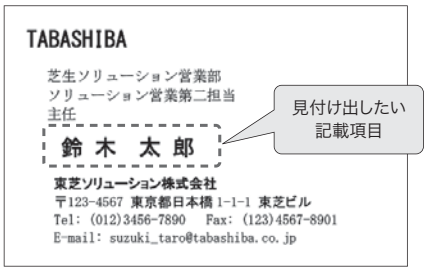
(記載内容)
字種:
0123456789

(b) 見出しによる記載項目の定義

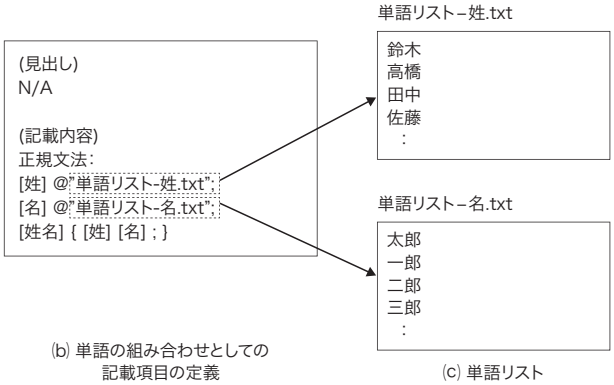
図3. 見出しで見付け出す記載項目

見出し文字列「お客様番号」で，文書中から当該記載項目を探し出すことができる。

Item found by its label



(a) 見付け出したい記載項目

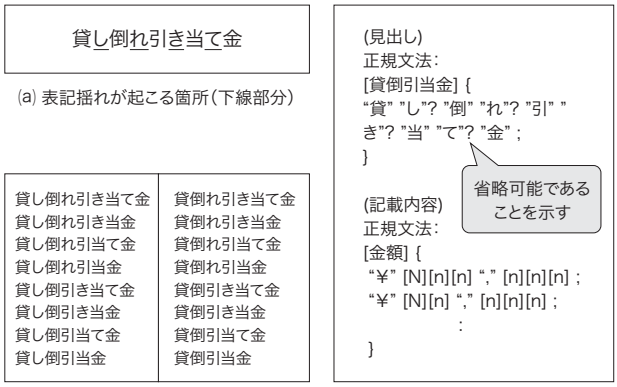


(b) 単語の組み合わせとしての記載項目の定義　(c) 単語リスト

図4. 記載項目自体の字面で見付け出す記載項目

記載項目をリスト中の単語の組み合わせとして定義し，定義に合致する文字列を見付け出す。

Item found by pattern of its character sequence



(a) 表記揺れが起こる箇所(下線部分)　(b) 表記揺れによるバリエーション　(c) 正規文法による記載項目の定義

図5. 表記揺れを含む見出しを定義する例

正規文法で見出し文字列を記述することで，4か所ある送り仮名の有無の組み合わせ$2^4=16$通りを簡潔に定義できる。

Example of label defined by regular grammar

(3) 表記揺れを伴う見出し文字列を定義する例　見出しには，送り仮名の有無などによる表記揺れを伴うことも多く，例えば**図5**に示した例では4か所の送り仮名の有無により，$2^4=16$通りのバリエーションが想定される。このような表記揺れの問題に対しては，単語リストへのバリエーションの登録による対応も考えられるが，この技術では，図5(c)に示すような正規文法で，より簡潔に表記揺れを含む表現を定義できる。

このように正規文法による読み取り項目の定義にフレキシブルOCR知識処理を適用することで，住所，姓名，電話番号，日付などの典型的な読み取り対象に加えて，会社名などの複雑な読み取り対象もパターン化でき，表記揺れにも効率良く対応できる。

4. 表の認識

項目サーチOCRでは，表の認識と内容の論理的な解釈をサポートしている。**図6**で例示した帳票の表には，上側に見出しが並んでいて，その下に対応する値が複数，縦に並んでいる。

表を含む帳票の認識結果の例を**図7**に示す。この例では，行ごとに，記載項目固有の識別名とそれに対応する値が並んでおり，例えば識別名「06_資本金」で始まる行を抽出することで資本金の値を取得できる。

このように項目サーチOCR技術では，記載内容の論理的な解釈が認識結果に反映されており，後段のアプリケーションやRPAソフトウェアで認識結果を活用できる。

株主資本等変動計算書

（単位：円）

有限会社 赤字屋食堂　　（自 平成26年 3月 1日 至 平成27年 2月28日）

	資本金	資本剰余金	利益剰余金	純資産合計
前期末残高	5,000,000	0	△265	4,999,735
当期変動額				
剰余金の配当額	0	0	0	0
当期純利益	0	0	△1,790,904	△1,790,904
当期変動額　合計	0	0	△1,790,904	△1,790,904
当期末残高	5,000,000	0	△1,791,169	3,208,831

図6. 表を含む帳票の例

上側に見出しが並んでいて，その下に対応する値が複数，縦に並んでいる。

Example of sheet containing table

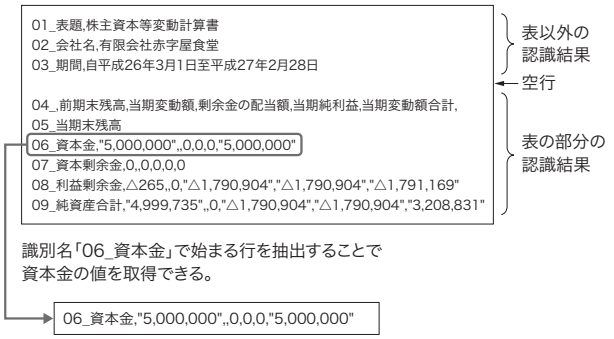


図7. 表を含む帳票の認識結果の例

表の論理的な解釈が認識結果に反映されており，識別名で表の内容を認識結果から抽出することができる。

Example of result of recognition of sheet containing table

5. あとがき

ここでは，読み取り対象を文書中から自動的に見付け出す項目サーチOCRについて述べた。この技術を用いることで，モデルによる座標指定なしで多様な様式や，レイアウトや，読み取り項目の表記揺れに対応できる。また，表形式で記載された項目についても，アプリケーションやRPAソフトウェアと連携できるように，論理的な解釈が反映された認識結果を出力できる。

今後は，認識対象や見出しの文字列を自動的に学習し，多種多様な帳票に効率的に対応できる仕組みの実現を目指す。

文　献

(1) 宇田明弘，ほか．“表形式既存帳票認識システム”．第1回情報科学技術フォーラム一般講演論文集第3分冊．東京，2002-09，情報処理学会，2002，p.167–168.

(2) Lam, S. W. et al. "Anatomy of a Form Reader". ICDAR (International Conference on Document Analysis and Recognition) '93, Ibaraki, 1993-10, ICDAR, 1993, p.506–509.

(3) 平山淳一，ほか．仮説検証型アプローチを用いた定義レス非定型帳票認識技術．電子情報通信学会論文誌D．2014，J97-D，12，p.1797–1808.

(4) 鈴木智久，中島康裕．決定性有限オートマトンと文字候補ラティスの照合によるフレキシブルOCR知識処理．東芝レビュー．2015，70，4，p.46–49.

鈴木　智久　SUZUKI Tomohisa
東芝デジタルソリューションズ（株）
ソフトウェア& AI テクノロジーセンター 知識・メディア処理技術開発部
電子情報通信学会会員
Toshiba Digital Solutions Corp.

横田　和章　YOKOTA Kazuaki, Ph.D.
東芝デジタルソリューションズ（株）
ICT インフラサービスセンター ソフトウェア開発部
博士（工学）　電子情報通信学会会員
Toshiba Digital Solutions Corp.