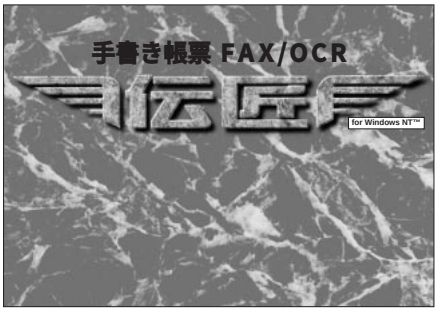

リコー文書画像認識SDK

RICOH Document Image Recognition Software Development Kit

松下　貢*　鈴木　俊博**　内山　幸央**
Mitsugu MATSUSHITA　Toshihiro SUZUKI　Yukinaka UCHIYAMA
別所　吾朗**　長谷川　史裕**
Goroh BESSHO　Fumihiro HASEGAWA

要　旨

リコー文書画像認識SDKは，紙を媒体とする情報をコンピューター上で扱えるようデータ変換するために，文字認識他の機能をもったWindows95/NT上のソフトウェアツール群である．主な特徴は以下のとおりである．

1) 画像補正，文字認識など文書画像認識に必要な様々な機能をもつ
2) 簡単かつ柔軟な文書画像認識システムの構築が可能
3) 専用ハードウェアを必要とせず，手書き漢字認識を実現
4) OCR専用帳票でなく一般流通帳票の認識が可能

ABSTRACT

A Ricoh Document Image Recognition SDK is the name of software tools of windows 95/NT including character-recognition, etc. which convert paperbased information to data easy to handle in computer.

1) Many functions for the document image recognition system, including enhancement of scanned image, character recognition, and also.
2) Easy and flexible implementation for document image recognition system.
3) Handwritten Kanji character recognition without a specific hardware.
4) Recognition of a document with a non standard form for OCR.

1．背景と目的

日本においてOCR(光学的文字認識装置)の製品化が始まったのは1968年前後である。OCRのその後の進歩はめざましく、OCRでの処理対象は英数字、カナ、記号から漢字へ、活字から手書きへ、専用帳票から一般文書(一般帳票)へと広がってきた。製品形態も専用ハードウェアによるセンター集中処理型からクライアント側での分散処理型へと変化してきた。こうした製品化の歴史の中でOCRをとりまく環境は大きく変化した。オープン化、ソフト化、部品化である。このような環境の変化の中でリコーは一貫して文字単体の認識だけでなく、紙を媒体とする情報を全体としてコンピューター上で扱えるようデータ変換すること、オープンな環境の中でソフトウェア部品として実現できることを指向して研究開発を行ってきた。こうした研究開発の成果を受けて、システムソリューション事業本部で、文書画像認識に必要なソフトウェア部品としてWindows95/NT上で利用できるライブラリー(DLL)やOCXを開発し、「リコー文書画像認識SDK」の総称のもとにソフトウェア開発キット(SDK)の商品化を進めている。本稿では、文書画像認識SDKの概要について、特に受発注書などの定型帳票を認識するために必要なライブラリーを中心に述べる。

* システムソリューション事業本部　ドキュメントシステム開発部
Document System Development Department,
System Solution Business Division
** 研究開発本部　情報通信研究所
Information and Communication R&D Center,
Research and Development Group

2．製品の概要

2-1　文書画像認識SDK

文書画像認識SDKとはOCRをはじめとする文書画像認識技術の各技術をライブラリー(DLL)やOCXなどのソフトウェアの部品としてユーザーに提供するために商品

化した商品群の総称のことである。この文書画像認識SDKには次のような商品が含まれている。

・活字OCR(印刷された文字を認識)
・手書きOCR(手書きで書かれた文字を認識)
・定型帳票処理(定型帳票に対して文字領域を抽出)
・OCR用知識処理
(住所、氏名などの知識情報でOCR結果を修正)
・スキュー補正(傾いた画像を傾きのないように補正)
・領域識別
(タイトル、表など文書のレイアウトを自動解析)
・表処理(表中の文字領域や罫線情報を抽出)
・ノイズ・網掛け除去
(ノイズや文字にかかった網掛けを除去)
・タイミングマーク検出
(帳票処理で使われるタイミングマークを検出)

これらのライブラリーはすべてソフトウェアで実現されており、Visual C++などのソフトウェアツールを用いてユーザーがシステムに簡単に組み込むことが可能である。また、各ライブラリー間での共通部分(画像に関する構造体など)の共有化なども取られており、複数のライブラリーの組み合わせも簡単に行うことができる。

文書画像認識SDKは、リコーで開発、商品化されている活字OCRソフト「読取物語」やFAX/OCRソフト「伝匠」などにも利用されているので、「読取物語」での利用例、「伝匠」などの定型帳票OCRでの利用例をFig.1のフローで説明する。「読取物語」では、傾いた画像を補正するスキュー補正、ノイズ、網掛け文字の網を除去するノイズ・網掛け除去、レイアウト解析を行う領域識別、表中の文字や罫線を抽出する表処理、活字文字を認識する活字OCRの各SDKを組み合わせて商品化を行っている。定型帳票OCRでは、スキュー補正、ノイズ・網掛け除去、帳票中から文字記入領域を抽出する定型帳票処理、手書き文字を認識する手書きOCR、OCR認識結果を住所・氏名などの知識辞書を用いて修正するOCR用知識処理などのSDKを利用すれば良い。

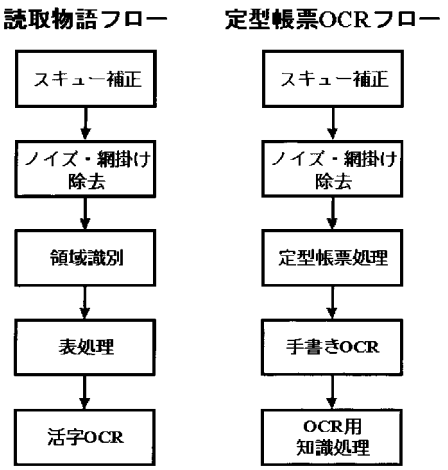

Fig.1

2-2 定型帳票OCR

ここでは、文書画像認識SDKの中から定型帳票OCRの開発に必要なライブラリーについて述べる。

定型帳票OCRとは、受発注書など常にフォーマットの決まっている帳票を大量に入力しなければならない場合にその帳票を自動的に処理するために、あらかじめ、その帳票のフォーマットをシステムに登録しておき、そのフォーマット情報を基に文字記入領域の文字画像の抽出を行った後、その文字画像をOCR処理するシステムのことである。ここでは、まず定型帳票OCRの流れを述べる。

0) マスター帳票登録

定型帳票処理を行う場合には、事前にその帳票のフォーマットを登録しておく必要がある。具体的には、文字が記入される領域座標の指定、その領域で認識する文字の文字種(数字／英字／漢字など)、知識処理を行う場合にはその知識処理の種類(住所／氏名／商品名など)などを登録する。なお、この処理は新規の帳票を登録する際に一度だけ行えば良いものであり、通常の認識処理を行う際には必要ない。

1) 帳票入力＋画像前処理

これ以降が実際に文字の記入された帳票を認識する過程である。まず、文字が記入された帳票をスキャナーまたはFAXで画像として取り込む。この画像を以後の処理で認識しやすい画像にするために何らかの前処理を行うのが一般的である。具体的には、帳票画像に傾きがある場合にはスキュー補正、FAX画像のようにノイズが多い画像に対してはノイズ除去などを行う。

2) 文字領域抽出(定型帳票処理)

入力された帳票画像から文字記入領域を抽出する。

3) OCR(手書きOCR)

定型帳票処理で抽出された文字画像の認識を高速、高精度に行う。

4) OCR認識結果自動修正(OCR用知識処理)

OCRで得られた認識結果を住所、氏名などの知識辞書を用いて修正する。

以下、定型帳票OCRの中でキーとなる技術、定型帳票処理、手書きOCR、OCR用知識処理について以下に述べる。

2-2-1 定型帳票処理

入力された帳票画像の中からあらかじめ登録しておいたフォーマット情報を基に文字記入領域に書かれた文字画像を抽出する。

定型帳票処理の機能・特徴：

・従来のOCR専用帳票のようなタイミングマーク、ドロップアウトカラーによる枠線が不要
→既存の流通帳票やワープロで作成した帳票の利用が可能

・マスター帳票画像と入力画像との位置ズレに強い
・文字と文字記入枠が接触しても文字抽出可能
・マスター帳票の登録も簡単

2-2-2　手書きOCR

手書きで書かれた文字の画像を認識し、テキストとして出力する。

手書きOCRの機能・特徴：

・漢字までを含む多文字種の手書き文字の認識をソフトウェアで高速、高精度に実現
・認識可能な文字種：数字、英字、カタカナ、ひらがな、記号、漢字(漢字は第1水準＋第2水準約500文字の約3500文字)
・数字については特に低誤認識を実現

2-2-3　OCR用知識処理

OCRでの認識結果誤りを住所、氏名などの知識辞書を使って自動修正する。

OCR用知識処理の機能・特徴

・住所：日本全国の住所の地域名まで登録(約11万件)
・氏名：頻度の高い姓、名を登録(約9,000件)
・商品名、商品コードなどのユーザ登録が可能
・住所と郵便番号、商品名と商品コードなどの関連付けが可能

3．技術の特徴

3-1　定型帳票処理

従来の定型帳票OCRシステムではOCR専用帳票というOCR処理しやすい専用帳票を使用するシステムが多かった。OCR専用帳票では、帳票の位置合わせを正確に行うためにFig.2のようにタイミングマークというマークを付加したり、文字記入枠と文字との接触を避けるために文字記入枠をドロップアウトカラーという画像として取り込んだ際に線が残らないような色で印刷していた。このようなOCR専用帳票を利用するには、印刷所に帳票の作成を依頼しなければならないので、帳票作成にかかるコストが高かった。また、従来のシステムでは、

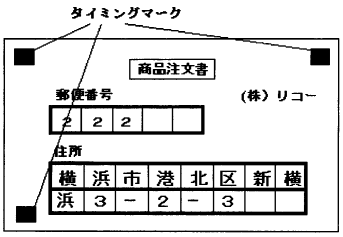

Fig.2

自分でワープロを使って作成した帳票や既に使用しているOCRを意識していない帳票を処理することもできなかった。そこで、リコーでは、このような従来利用することができなった帳票でも利用することが可能な定型帳票処理技術の開発を行った。

定型帳票処理技術は大きく分けて画像位置合わせと文字画像抽出の二つの機能に分けることができる。画像位置合わせ機能とはあらかじめ登録しておいたマスター帳票と実際に入力される帳票との画像の位置ズレを自動的に補正する機能である。一般に画像は入力される条件が毎回異なるため、Fig.3のように位置ズレが生じるのでこの位置ズレを自動的に補正する。文字画像抽出機能とはFig.4のように文字記入領域から文字を自動的に抽出する機能である。

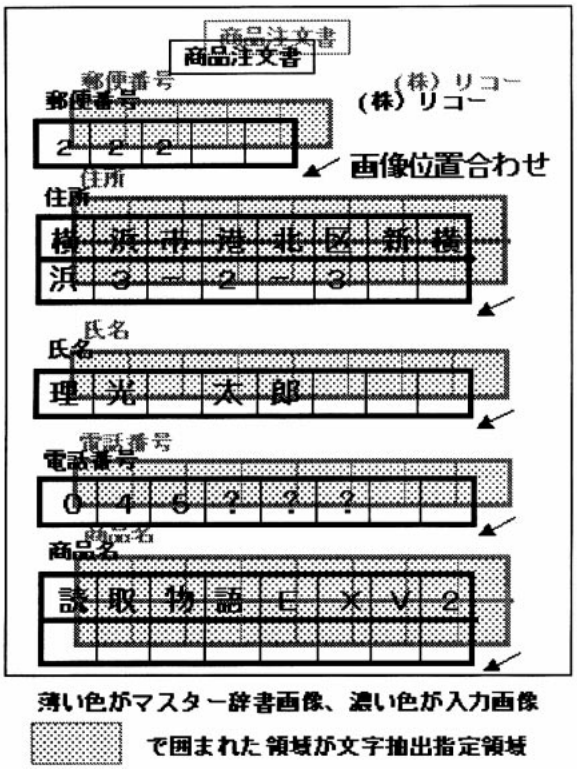

Fig.3

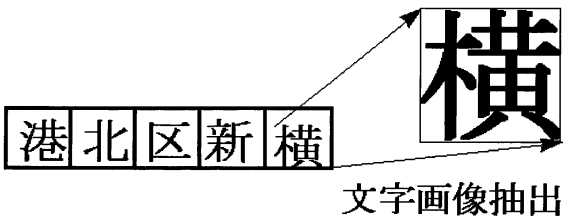

Fig.4

画像位置合わせ機能では、従来のシステムで使われているようなタイミングマークを必要とせず、帳票中レイアウト情報(デザイン情報)を利用して、マスター帳票画像と入力帳票画像との位置合わせを行う。この位置合わせ機能では、スキャナーやFAXの機種の違いなどによる若干の画像伸縮にも対応している。例えば、 FAXの

場合、解像度はファインモードで約200DPIであるが、正確には主走査方向204DPI、副走査方向196DPIであるため、スキャナーで取り込んだ200DPIの画像とは若干の大きさの違いが生じるが、このような問題にも対応することが可能となる。

文字画像抽出機能は、画像位置合わせ機能でマスター画像との位置合わせを行った画像に対して、あらかじめ指定しておいた文字記入領域中の文字画像を抽出する。その際に不要な罫線を消去するので、従来のシステムでは困難であったFig.4にある「横」という文字のように文字と文字記入枠との接触があった場合でも正確に文字を抽出することが可能となった。

また、この定型帳票処理技術を用いれば、OMR(マーク認識)の開発も簡単に行えるようになる。

3-2 手書きOCR

手書きOCRでは、数字、英字、カタカナ、記号だけでなく、ひらがな、漢字までの3500種以上の文字の認識をソフトウェアで実現している。手書きで書かれた文字は印刷された活字と比べて、個人差によるばらつきも大きく認識が困難なため、大量の情報を利用したり、複雑な処理を行う必要があるので、従来は専用のハードウェアを使ったシステムが多く、ソフトウェアで実現されているシステムはまだ数少ない。

OCRで文字を認識するには、まず文字画像の中から認識の手がかりとなるような特徴的な要素を抽出する。ところが手書き文字の場合、個人によって特徴的な要素がなかったり、反対に強調されすぎていたりするので、多様な観点から特徴要素を抽出しなければならない。こうして扱うデータが増えると、計算量が増え十分な認識速度が達成できなくなるほか、認識用の辞書ファイルが巨大になってしまう問題も発生する。リコーではこのようにソフトウェアで実現することが困難であった手書きOCRを、効果的な情報圧縮手法や3500を越える文字の中からあらかじめ簡単な処理で少ない候補に文字数を絞り込む大分類技術の開発により実現可能にした。

商品の注文書において数字は、仕入れの数量や金額として使われる。このような項目で誤認識が発生した場合、致命的な問題が起きるので、他の文字種に比べて誤認識による影響が大きい。しかも、数量や金額で記入される数字は文字と文字のつながりの規則性がないので、3-3で述べるOCR用知識処理を利用した認識結果自動修正ができない。自動修正ができず認識結果に誤りがあった場合、その修正は人が目視でチェックし、手入力で修正することになる。その際、認識結果がリジェクト(認識結果に自信がないので分からないという結果)であれば、修正箇所を簡単に見つけることができるが、誤認識の場合、修正箇所を見落とす確率も高くなる。したがって、このような数字の認識の場合には、さらに詳細な解析処理を行い、まぎらわしい書き方であったり、文字画像が劣化している場合など、確実な認識が期待できない文字画像に対してはリジェクトを返すようにし、なるべく誤認識が少なくなるよう工夫している。

この他にも金額欄などで数字を記入する場合、Fig.5のように連続する「0」という文字を連結して記入することが多いが、このような他の文字と連結した数字の認識にも対応するなどユーザーが利用しやすい工夫を行っている。

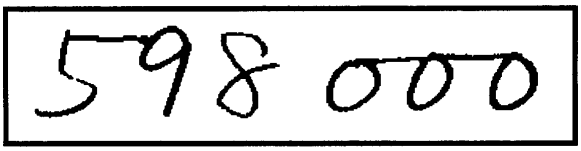

Fig.5

3-3 OCR用知識処理

3-2で手書きOCRの説明を行ったが、数字の認識精度は高く、認識率は99%以上である。しかし、漢字までの文字種を含む手書きOCRの認識精度はまだまだ十分な精度が得られておらず、当社データベースでの測定結果でも認識率90%程度というのが現状である。しかし、定型帳票で手書き文字を認識させたい場合には、ある程度記入される文字列が特定できるので、このような情報を知識として与えることにより認識結果の自動修正が可能になる。定型帳票で入力する項目としては、住所、氏名、商品名、商品コード、個数、金額などが多いと思われる。このうち、住所、氏名、商品名、商品コードについては、あらかじめ用意した単語情報との照合により認識結果を修正することができる。例えば、住所であれば全国の地名、氏名であれば姓や名、商品名であればすべての商品名を登録した辞書を用意しておく。現在のOCR用知識処理ライブラリーでは、日本全国の地域名までの辞書(約110,000件)と人名(姓、名)の辞書(約9,000件)を提供している。また、商品名などユーザー側で設定したい項目がある場合には、ユーザーが新規に辞書を作成／追加できるようになっている。

以下に、OCR用知識処理の実際の処理例を示す。

Fig.6 のように「横浜市港北区新横浜」を「横洪市港此区新横洪」と誤認識したとする。OCR認識結果の下に書いてある文字はOCRで2位以下の候補にあげられた文字である。知識処理では、このOCR認識結果と候補文字を基に住所に関する知識辞書を使ってOCR結果の修正を行う。まず、最初の「横浜市」の部分に着目してみると、候補文字を利用して「横洪市」、「積洪市」、「接洪市」、「横浜市」などの組み合わせを調べていき、実際に存在す

る「横浜市」を選択する。また、正解の文字が候補文字に含まれていない場合であっても、「横洪市港此区新横洪」というように住所に存在しない文字列であれば、それにもっとも近い「横浜市港北区新横浜」に自動修正することができる。同時に、「横浜市港北区新横浜」という住所は郵便番号「222」と特定できるので、郵便番号の認識結果も合わせて修正することができる。

このOCR用知識処理を用いることにより、当社データベースにおいて、手書きOCR単独での認識率約90%を約96%まで高めることができた。

4．今後の展開

リコー文書画像認識SDKはソフトウェアの部品であるため、ユーザーがシステムに組み込んだ後、バージョンアップを行う際に必要なソフトウェアの部品のみを入れ替えることにより機能、性能の向上ができるというのも一つの利点である。よって、現在商品化している各技術のさらなる精度向上は常に進めていかなければならない。また、ユーザーが必要とする技術がこの他に見つかれば、それらの技術を随時開発し商品群に加えていくことにより、より魅力のある商品として育てていきたい。

また、今回は定型帳票OCRを中心とした文書画像認識SDKを紹介したが、システムソリューション事業本部ではこの他に、音声合成、音声認識の機能を揃えた音声SDK、全文検索、日本語解析の機能を揃えた日本語SDKなどの商品化も行っているので、これらの商品との融合やこれらの技術を利用したシステム／アプリケーションソフト開発にも注力していきたい。

最後に本ソフトウェアの開発にあたり多くの方々にご指導、ご協力いただきましたことを深く感謝致します。

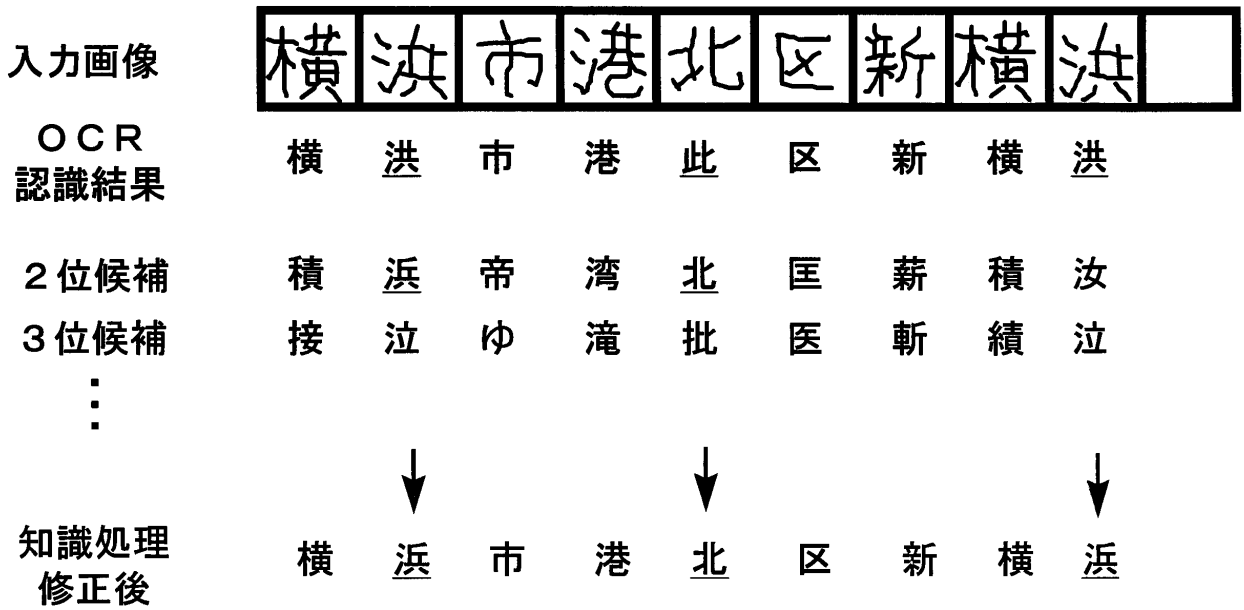

Fig.6