デジタル・ドキュメント　27－5
（２００１．３．１４）

イメージドキュメント処理方法の一検討

後藤　裕久　　政井　宏之　　樋口　浩一

沖電気工業株式会社　情報技術開発センタ

紙資料に記載された文字を情報システムに入力するために、文字認識技術を用いたＯＣＲシステムが盛んに使用されており、近年でも出荷台数、金額ともに増加傾向にある。ＯＣＲのシステム形態として、ネットワーク経由でＯＣＲサーバと確認修正を行うクライアントが接続される形態が開発されている。最近のＷｅｂシステムの急速な普及で、クライアント側に専用アプリケーションソフトを載せる形態ではなく、クライアントの環境を軽くするためにブラウザベースで操作を行えるニーズが出てきた。本稿では、ＸＭＬを利用して文書イメージの連携を取りながら文字認識結果の確認および修正を行えるイメージドキュメント処理方法を検討した。

A Study on The Image Document Processing

GOTO Hirohisa, MASAI Hiroyuki, HIGUCHI Koichi

OKI Electric Industry Co., ltd.　Information Technology R&D Division

The OCR system which used character recognition technology is very much being used, and it shows a tendency of increasing even recently in shipping number, an amount of money to input the characters in the paper to the information systems. In the OCR systems the original data format, including OCR results and images, is used between the client and the OCR server. This paper describes Web-based viewer and editor system of the OCR result data with XML.

1．はじめに

近年のブラウザ等インターネットを使用して個人ベースでデータを入力するシステムは、急速に広がっている。しかしながら、現状は、紙の帳票に必要事項を記載して、窓口やセンタに提出し、その紙帳票からデータを入力する形態が多く行われている。たとえば金融機関では個人の申込書など窓口で紙の帳票を用いており、それらの帳票のデータ入力は人手によるパンチ入力やＯＣＲによって入力で行われている。すなわち、顧客は紙帳票にデータを記入し、情報システムに入力するために紙帳票からデータを抽出するという形態である。個人による分散した入力形態ではこのような紙帳票を使用したデータの受け渡しが容易であり、データの確認が容易に行われる。そして、紙帳票から情報システムへのデータ入力として人手によるパンチ入力よりも高効率のためＯＣＲの導入が進められている。

まず、ＯＣＲの処理手順を図１「ＯＣＲ処理概要」のブロックに対応して以下説明する。

（1） 画像処理

スキャナもしくはＦＡＸなどを用いて入力された文書画像に対して文字認識に適した画像処理を行う。たとえば、画像２値化処理、文書の位置検出、文書画像の傾き検出／補正、画像の回転などの処理を行う。

（2） 帳票レイアウト識別

入力文書画像からレイアウトを抽出する。

ＯＣＲ専用帳票の場合には、帳票につけられたシートＩＤを識別して帳票のレイアウト情報を取得する。レイアウト情報はあらかじめシートＩＤに対応して定められている。

ＩＤのない帳票では、文書画像から画像の特徴を抽出して、レイアウト情報を取得する。

（3） 文字認識

帳票レイアウト識別で得られたレイアウト情報にに基づいて、文字認識を行う領域の中から文字画像を切り出して文字認識を行う。このときに、文字が含まれる領域の座標、行数、行内の文字数などの情報も抽出する。

（4） 認識後処理

文字認識で得られた文字コードは一文字単位の認識結果のため、対象とする文字列の規則で認識結果の修正を行う。

たとえば、住所が記載されている文字領域であれば、住所の知識辞書との整合をとり、認識結果の修正をおこなう。

（5） 出力データ編集

上記の処理で得られた文字コードとその座標、文字が含まれる行の座標、入力画像などをまとめて、認識結果の確認・修正を行うために必要なデータに編集する。

（6） 認識結果確認修正

編集された出力データを人間の目で見て、正しく認識されているかを確認し、誤りがあれば修正する。画面に表示された認識結果を目視で確認し、キーボードなどを用いて修正する。

これらＯＣＲによるデータ入力の工程（1）から（6）の中で、（1）から（5）はコンピュータよる自動処理であるが、工程（6）は人手による確認および修正作業が行われる。そのため、多くの帳票を処理するデータエントリ部門においては、複数の人間による確認および修正作業を行う。確認および修正を行うシステムでは、データ転送の高速性などからデータ形式は固有の形式を用いられる。

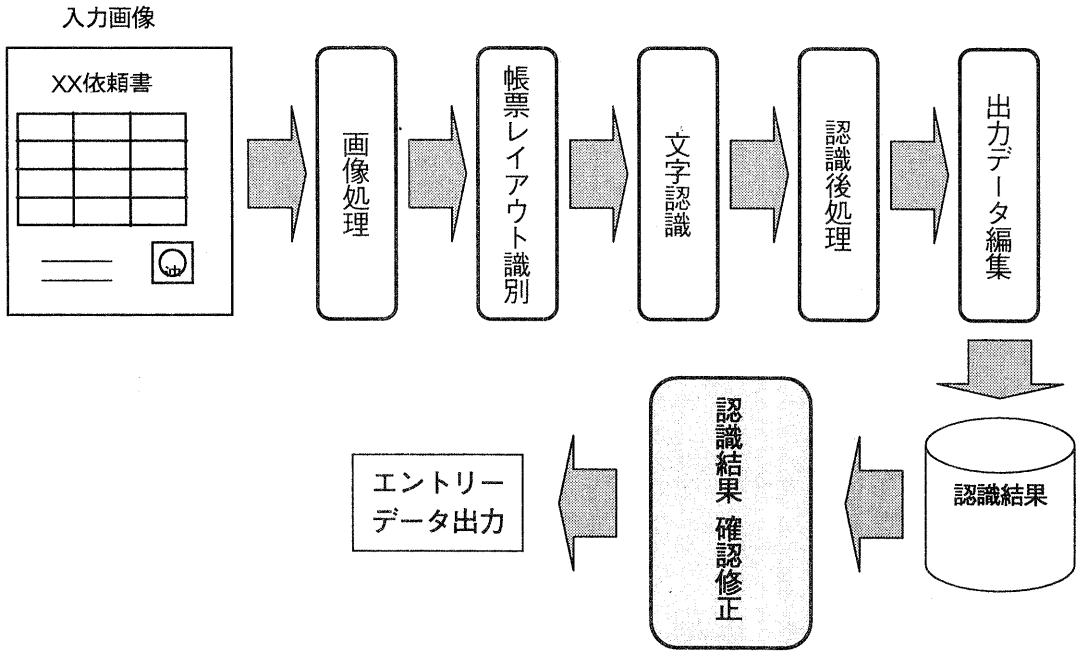

図1　ＯＣＲ処理概要

本稿では、固有のデータ形式を扱う専用アプリケーションソフトをクライアントに載せる形態から、クライアントの環境を軽くするためにブラウザベースで操作を行う形態を検討するために、認識結果をＸＭＬで記述し、インターネットを経由して分散した環境で認識結果の確認および修正を行うシステムについて検討した。

2．確認および修正のためのデータ形式

認識結果の確認および修正で使用するデータには下記のものがある。

・文字認識結果：　文字画像を認識してえられた文字コードと文字認識候補

一般に、漢字やひらがな、カタカナ、英字など文字種が多いものは文字認識において文字画像から得られた特徴に基づいて、予め定めた認識用辞書との類似性で判定を行うので類似した文字を候補として出力する。

認識結果に修正を行う場合に、あらためてキーボードから入力を行うのではなく、この文字認識候補から正解を選択して認識結果を入れ替えるという操作が容易である。

・画像表示：　文字認識結果を確認するために、文字認識結果に対応する文字画像もしくは文字画像を含む領域の画像を表示して、認識結果との比較を行う。さらに必要であれば入力文書の画像を見て確認を行う。

これは、文字認識を行う対象の文字画像を抽出する際に入力画像におけるノイズなどの影響で文字画像の切出しを誤る場合があり、認識結果だけを見ても修正が行えない。

そのため、認識結果に対応する文字画像を表示することで、確認および修正を行うオペレータの作業が確実に行える。認識結果に対応する文字画像の表示は、入力文書の全体画像もしくは部分領域の画像と文字枠などの座標を用いる。

```
<文書情報>
  <全体画像>
    文書画像ファイル名                          ・・・入力画像：入力文書全体の画像
  </全体画像>
  <領域情報>                                    ・・・認識対象となる領域の情報
    <領域画像>
      領域画像ファイル名                        ・・・認識対象となる領域の画像
    </領域画像>
      <領域座標>
        <開始X座標>  10  </開始X座標>                ・・・領域座標　矩形
        <開始Y座標>  20  </開始Y座標>
        <終了X座標>  100  </終了X座標>
        <終了X座標>  200  </終了X座標>
      </領域座標>
      <行数>  1  </行数>                            ・・・領域内部の行数
      <行情報>                                      ・・・領域内の行数分ある
        <文字数>  10  </文字数>                     ・・・行に含まれる文字数
        <文字情報>                                  ・・・1行の文字数分の情報
          <文字座標>
            <開始X座標>  15  </開始X座標>           ・・・文字領域座標　矩形
            <開始Y座標>  25  </開始Y座標>
            <終了X座標>  70  </終了X座標>
            <終了X座標>  80  </終了X座標>
          </文字座標>
          <文字結果>
            <候補　確信度＝２１０>  あ  <候補>      ・・・文字認識で得られた第1位候補
            <候補　確信度＝１８０>  め  <候補>      ・・・文字認識で得られた第2位候補
```

図2　文字認識結果のデータ形式例

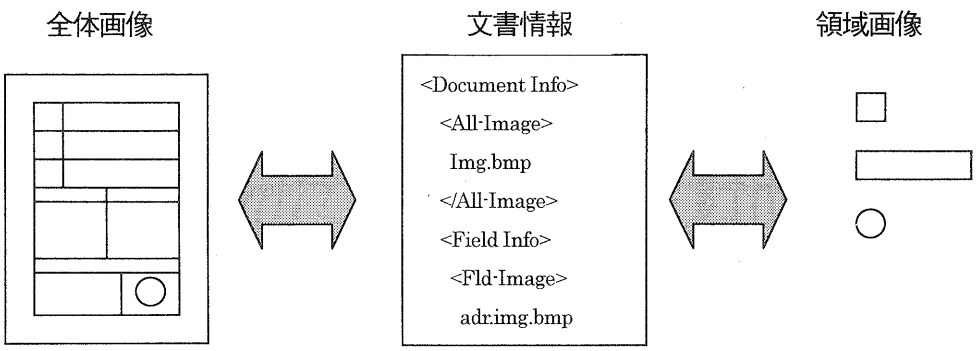

図3　文書情報と関連画像の連携

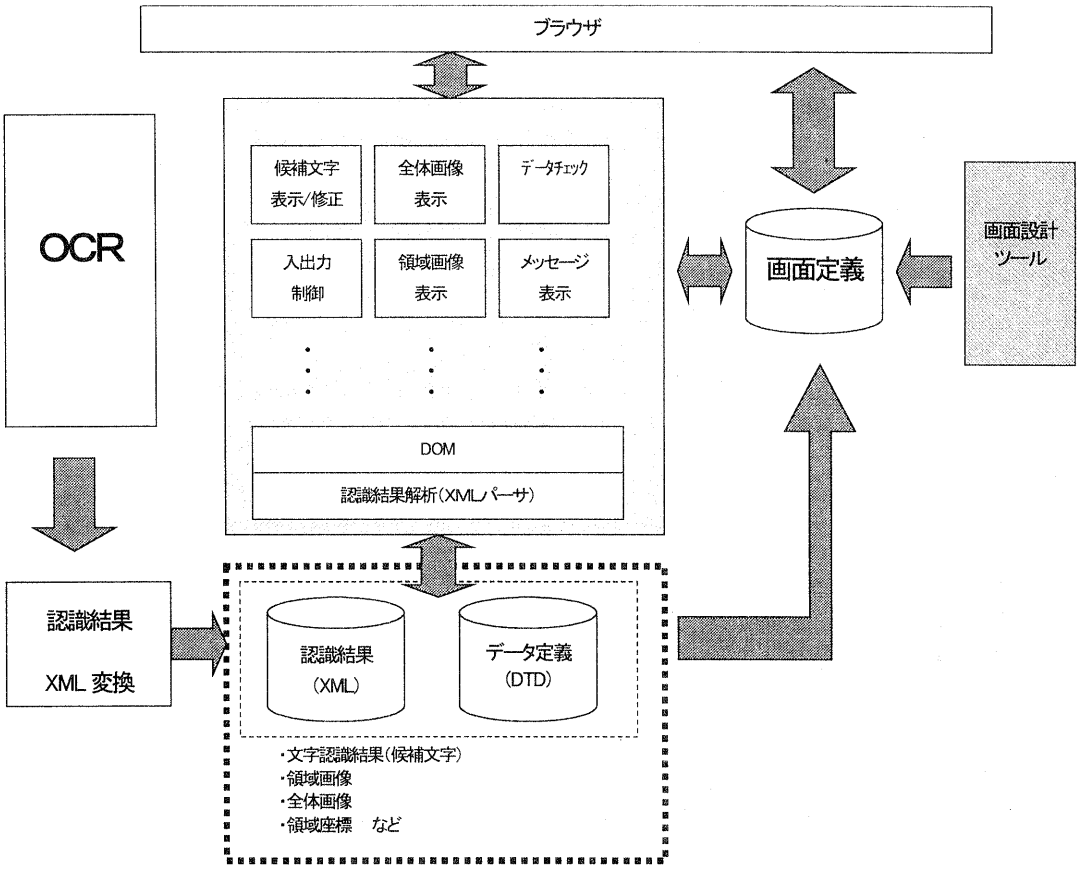

図4　文字認識結果修正確認システムの構成

上記の作業を行うために必要なデータの例を図2に示す。

ＯＣＲ出力から得るデータを分類すると、

（1）　エントリー情報

領域の番号などとその領域の文字コードなど、エントリーして出力されるデータ

（2）　確認および修正に必要な情報

文字認識候補、文字枠の座標、全体画像、領域画像などオペレータが認識結果を画面で確認・修正するために必要な情報

（3）　ＯＣＲ処理の結果情報

ＯＣＲ処理の工程で行った処理の履歴、たとえば、画像の傾き角度、画像2値化の閾値、画像の回転角度など。

があり、認識結果の確認および修正作業では、上記（2）で述べたように図3に示すような文字認識結果を含む文書情報と文書画像（全体画像と領域画像）と連携した画面表示が重要である。

3．実験

ＸＭＬのデータ形式で作成されたＯＣＲの文字認識結果をブラウザで確認修正する実験システムを図４のように構築した。従来のＯＣＲシステムに今回検討したブラウザベースの認識結果確認および修正システムを追加する形態である。

今回はブラウザベースの文字認識結果確認修正作業の実行を確認するためのために下記のような条件で行った。

・データ形式は、図２に示した形式
・ＸＭＬデータは、ＯＣＲ認識結果を参考にエディタで作成
・ＤＴＤを定義せず、タグ付けした認識結果ファイルを作成
・画面定義は、ＨＴＭＬで作成し、その上に機能を実現するコントロール実装

4．おわりに

本稿では、紙資料をスキャナやＦＡＸなどの入力装置で文書イメージにした後に、ＯＣＲで文字認識した認識結果をＷｅｂ経由で確認および修正するためのデータ形式としてＸＭＬを使用し、かつ、文字認識結果情報と文書イメージを連携させて扱うことで、容易なユーザインターフェイスを実現することが出来ることを述べた。

5．参考文献

・沖電気研究開発（No.156）『文字認識技術特集』（1992 年 10 月）