戸籍紙面の高精度認識方式[1)]

4D－10

岡本卓哉　樋野匡利　里　佳史
（株）日立製作所システム開発研究所

１．はじめに

これまで、紙による管理しか許されていなかった戸籍情報の電子化を認める法改正により、各自治体で戸籍情報システムの導入が検討されている。システム構築の際の重要な課題の一つに、これまで紙として管理されてきた情報の入力がある。この課題に対して、タイプ印字された戸籍紙面を認識し、コード情報に変換するシステムを開発した。本システムでは、戸籍の複数の書式を自動的に判定し、書式に応じて必要な情報を文字認識する。さらに、認識結果に対して戸籍の記述に関する知識を用いた後処理による検証と修正を行なうことで、高精度化を実現した。

２．戸籍認識システムの概要

本システムでは、図１に示したように、まず光ファイリング装置から戸籍の画像データを読み出す。読み出された画像に対して、順次、戸籍フォーマットの解析、文字認識、後処理による高精度化の処理を行なう。画像データと認識結果は、光ファイリング装置に出力して高速に印刷し、人手によるチェックと修正を行う。本システムは、WS(日立3050RX:PA-RISC,80MHz)と、文字認識装置(認識速度 100字/秒)を用いて構築した。

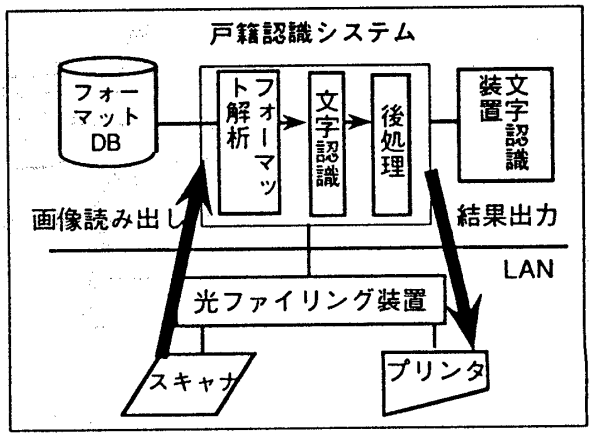


図１　戸籍認識システムの構成

３．フォーマット解析

戸籍には、本票、次票、附票など異なる種類のフォーマットが混在している。さらに作成時期によっては使用される用紙が違うため、同じ種類であってもフォーマットが多少異なる。したがって、入力された画像が複数のフォーマットのいずれであるかを自動的に判定することが必要となる。これを実現するために、戸籍のフォーマットをDBに登録し、順次、入力画像とのマッチングを行うことによりフォーマットを特定する方式を開発した。

本方式では、図２に示した手順で処理を行う。各処理の内容を以下に説明する。

（１）基準枠の抽出

図３に示すように、入力画像の周辺からサーチを始めて戸籍の罫線の外枠を抽出し、これを基準枠とする。この基準枠によって、画像の傾きや位置ずれを補正する。

（２）フォーマットマッチング

基準枠を基に画像上にDBのフォーマットを重ね合せる。次に、フォーマットを構成する罫線（フォーマット罫線と呼ぶ）により設定される範囲に含まれる罫線を画像から抽出する。抽出に成功した割合が最も高いフォーマットをマッチング結果とする。ただし、この割合が、あるしきい値以下の場合には、フォーマットエラーとし

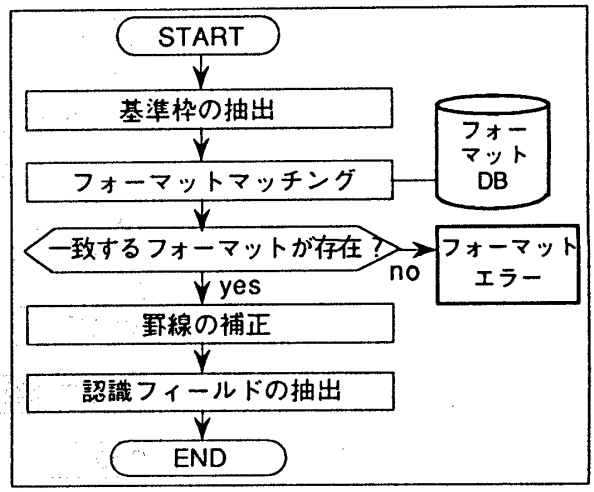


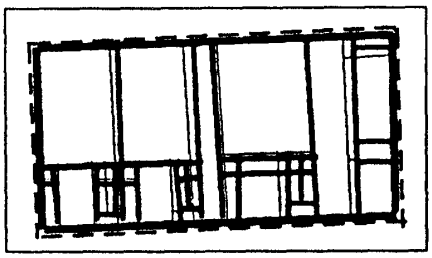
図２　フォーマット解析処理フロー

[1)] The Method of Family Register Recognition.
Takuya Okamoto, Masatoshi Hino, Yoshifumi Sato
Systems Development Laboratry, Hitachi,Ltd.

て処理する。

画像からの罫線抽出は以下の手順で行う。図4に示したように、まず画像上に重ね合せた各フォーマット罫線について、この罫線を挟んだ一定の幅の走査範囲を設定する。次に、この範囲において、フォーマット罫線と垂直な方向に画像を走査し、黒画素の連続（黒ラン）のうち、長さがしきい値より短いものだけを抽出する。抽出した黒ランの数が罫線の長さを基準としたしきい値より多ければ、罫線が存在すると判定する。

戸籍の入力画像



－－ 基準枠　── フォーマット罫線

図3　基準枠とフォーマット罫線

（3）罫線の補正

フォーマットマッチングで抽出された黒ランについて、その重心の点列に対する近似直線を求め、この直線をフォーマット罫線と画像とのずれを補正して得た罫線とする。

（4）認識フィールドの抽出

本籍などの認識対象となるフィールドは、フォーマットDB中に、そのフィールドの上下左右の罫線情報で記述されている。したがって、補正した罫線の位置情報を基に、罫線で囲まれた領域を認識フィールドとして抽出する。

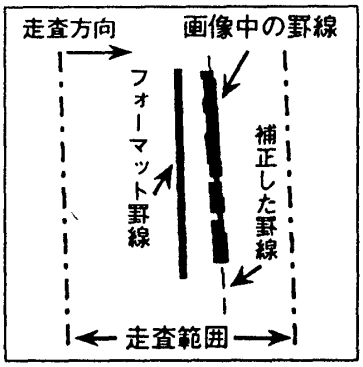


図4　罫線マッチング

4．文字認識および後処理

フォーマット解析で抽出された認識フィールドに対して、文字抽出と文字認識を行う。文字認識は、既開発の印刷文字認識方式[1]を適用し、認識結果として複数の候補文字およびその評価値を出力する。また、フィールドごとに出現する文字に制約がある場合は、認識辞書から対象文字のみを抽出して認識することにより、高精度化を図る。

認識結果に対しては、戸籍中に出現する定型句、単語などの知識を用いて後処理による検証と修正を行う。後処理では、図5に示すように、認識結果として出力される候補文字から、単語を抽出し、最適な単語の組み合わせを求める。戸籍認識のために登録した単語数は、地名を含めて約2,000単語である。年月日、地番については、値の範囲チェックなどによる検証、修正を行う。

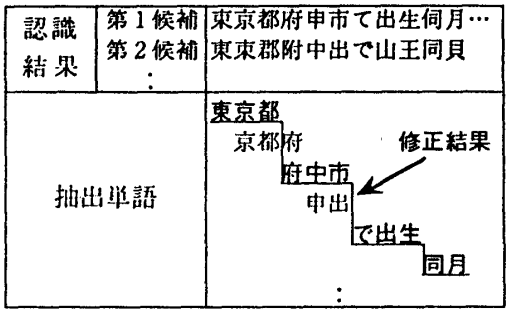


図5　文字認識の後処理

5．評価実験

評価用に作成した戸籍紙面100枚について認識実験を行なった。フォーマットはすべて正しく認識できた。文字認識率は表1に示す通りである。

氏名などの後処理による検証が行われない文字の認識率は約99%であったが、地名などの検証された文字に関しては、99.95%以上と高い認識率を実現した。また、本システムによる処理速度は、約200枚/時であった。

表1　文字認識率

総文字数	内訳			誤認識数(正解率)
29638	認識文字数	検証文字数	23965 (80.9%)	7 (99.97%)
		非検証文字数	4378 (14.7%)	48 (98.9%)
	リジェクト文字数		1295 (4.4%)	

6．まとめ

これまで開発してきた文書認識方式を応用し、フォーマット解析方式、文字認識の検証方式を戸籍向けに拡張することにより、高い認識率および認識速度を得ることができた。また、光ファイリング装置を利用することで、大量情報を処理する実用的なシステムが構築できた。

参考文献　(1)「黒画素方向性特徴のずらしマッチングによる印刷文字認識方式の開発」,岡本他,
情報処理学会第45回（平成4年後期）全国大会