2L－2

印刷文書認識システム AutoReco/2 －イメージプロセス－

平山唯樹† 山下晶夫† 加藤美治‡
日本アイ・ビー・エム株式会社 †東京基礎研究所/‡大和研究所

1 はじめに

印刷文書認識システム（AutoReco/2）では、対象となるイメージデータを取り込んだ後、イメージプロセスとしてまずレイアウト解析を行なう。レイアウト解析は文字列検出処理、図表分離処理、モデル照合の３つの処理からなり、これらの処理がシーケンシャルに進む。本稿ではこれらのレイアウト解析の概要を、図表分離処理に重点をおいて説明する。

また、文書画像のサンプルとして図1を用い、このイメージに対する処理の様子を示しながら説明する。

2 文字列検出処理

文字列検出処理の方法としては黒ランレングスの組合せによる手法を用いている [1]。この処理により、文書画像中の黒画素領域を囲む外接矩形が抽出される。これらの外接矩形は、あらかじめ与えたしきい値と高さとを比較することによって、文字列、縦罫線、横罫線、その他の黒画素矩形という４つの種類に分類される。サンプルから実際に抽出された黒画素矩形は図2に示す。

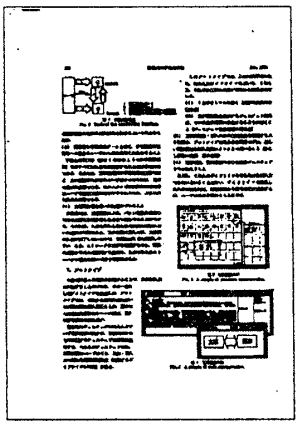

図 1: サンプルイメージ

Document Recognition System, AutoReco/2 –Image Process–
Yuki HIRAYAMA†, Akio YAMASHITA†, Yoshiharu KATO‡
†IBM Research, Tokyo Research Laboratory/ ‡Yamato Laboratory, IBM Japan Ltd.

3 図表分離処理

文書画像の構成要素としては、文字列領域と図表領域の大きく２つに分類することができる。しかし、これらを自動的に分類することは、（１）文字列領域だけでなく図表領域にも文字列が含まれる、（２）図表領域には、それ全体を囲っている矩形がないものも多い、（３）文字列領域と図表領域とが入り組んでいる場合もある、などの理由により難しい問題である。

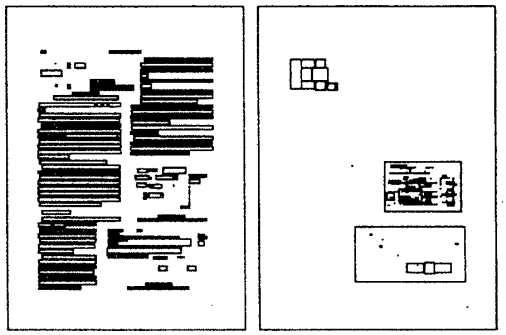

図 2: 文字列、縦横罫線、その他矩形

ここでは、文書画像中の文字列領域の規則性ならびに、文書全体のレイアウトの中心は文字列領域であるということに着目して以下のような処理を行なうことによって図表分離を実現している。図表分離処理では大きく分類して文字列の高さと距離の関係を用いたグルーピング、境界線を用いたページの分割、分割ブロックの統合という３つの処理を行なっている。これらは文字列検出処理によって得られた４種類の黒画素矩形をもとに処理を行なう。以下にそれぞれの処理について述べる。

3.1 文字列のグルーピング

文字列領域においては、同じような高さを持つ文字列が同じような距離を隔てて規則正しくならんでいる。これらの規則性を持つ文字列を以下の方法でグルーピングする。まず文字列矩形の高さとベースライン間の上下方向距離のヒストグラムを作る。それぞれのヒストグラム中の分布のまとまりを求める。さらに高さヒストグラムの中のまとまりごとの距離ヒストグラムを求め、

その距離ヒストグラムの最大のまとまりが全体の距離ヒストグラムの中のどのまとまりに対応しているかを調べる。この対応する距離ヒストグラム中のまとまりの最大値が、その高さヒストグラムのまとまりに属する文字列をグルーピングする際のしきい値になる。ある文字列が上下の文字列との距離がこのしきい値以内であれば一つのグループにグルーピングされることになる。この方法では、グルーピングのしきい値をページの情報からその文字列の高さに応じて求めることになる。

3.2 境界線を用いたページの分割

グルーピング処理によって、規則性を持ってならんでいる文字列は一つの大きなグループにまとめられる。これらのグループはページのなかでカラムを構成しているものもあり、全体のレイアウトを構成する大きな要素になっている。そこで、これらのグループからカラムの境界を求めるために、上下にならんでいるグループの左右の境界線の位置が一致する場合には境界線を一つにまとめると言う操作を行なう。そして、これらの境界線を上下に延長し、さらに境界線の上端下端で左右方向の境界線も作成する。そして、これらの境界線により図3に示すようにページ全体を分割する。

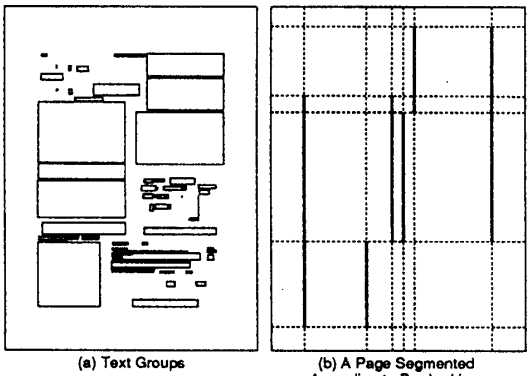

図 3: グループ、境界線

3.3 分割ブロックの統合

グループの境界線を利用してページを分割したが、このままではページ全体を細かく分割し過ぎる場合がある。つまり、ある部分を分割する境界線が他のひとまとまりになるはずの領域を分割してしまうことがある。そこで、複数の領域にまたがる黒画素矩形が存在する場合にはそれらの領域をひとまとまりにする。最後に以上の方法で求まった領域を、非文字列矩形の存在により文字領域か図表領域かを分類する。結果を図4に示す。

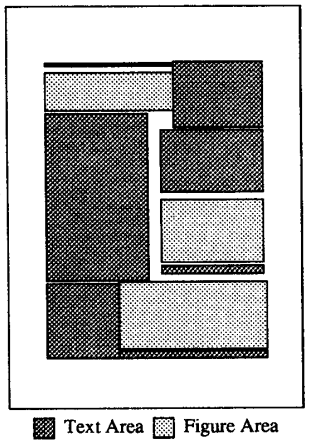

図 4: 最終結果

以上の方法を61ページ分の色々な分野からサンプルしたイメージに適用したところ、文字領域の93.3%、図表領域の93.2%を正しく抽出した。

4 モデル照合

イメージプロセスの最終段階として、レイアウトモデルを用いたモデル照合を行なう[2]。これは、部分–全体関係に基づく木構造によって表現されているレイアウトモデルをあらかじめ用意しておき、文書画像中の黒画素矩形をフィールドセパレータにもとづいて大きく分類した後、その黒画素矩形とレイアウトモデルを擦り合わせることによって実現されている。これによって、文書画像中の黒画素矩形に「タイトル」などのラベルづけが自動的にされることになる。

5 おわりに

あらかじめ自動的に文字領域と図表領域を分離すること、さらに領域に自動的にラベル付けすることは、文字認識システムを効率良く動かし、さらには大量文書の処理、無人運転を目指す上で非常に重要な機能である。今後はさらに研究を進めてこれらの機能を高めていく予定である。

参考文献

[1] T. Amano et al.: "DRS:A Workstation-Based Document Recognition System for Text Entry," IEEE Computer, July 1992, pp.67-71.

[2] 山下, 天野:" モデルに基づいた文書画像のレイアウト理解," 信学論, vol.J75-D-II, no.10, pp1673-1681, 1992.