５Ｖ－０３　　レイアウト変動の大きい文書画像からの項目自動抽出

岡田　康裕　　依田　文夫
三菱電機株式会社 情報技術総合研究所

１．はじめに

文書のレイアウトと記述内容双方の情報を用いて、レイアウト変動の大きい紙文書から文書項目を自動的に検出し、分類する方式について検討した。

２．全体構成

図１に処理の概要を示す。まず、文書画像解析により文字列の組合せで構成される文字列ブロックを抽出し文字認識を行った後、各文字列ブロックに対して文書のレイアウトに関する特徴と文書のコンテンツ（記述内容）に関する特徴を抽出する。次に、抽出した特徴と文書の各項目との照合を行う。照合はあらかじめ作成したレイアウトテンプレートと、ユーザがあらかじめ指定したキーワードリストを用いて、文書の各項目と文字列ブロックとの関連性を示す距離を求める。キーワードリストとの照合時には文字認識誤りが生じた場合でも正しくキーワード照合できるように、文字形状を反映した形状特徴を保持し、検索時には文字コード照合と形状特徴照合を併用してキーワード検索を行う方式を用いる。次に求めた距離値に対してしきい値処理を行うことにより、文書の各項目に対応する候補ブロックを抽出する。最後に各候補ブロックの距離値を初期確率として弛緩整合法により、候補ブロックの最適な組合せを決定する。弛緩整合法で用いる各項目の相対的整合性を表す適合係数は、項目間の相対的な位置関係により求める。

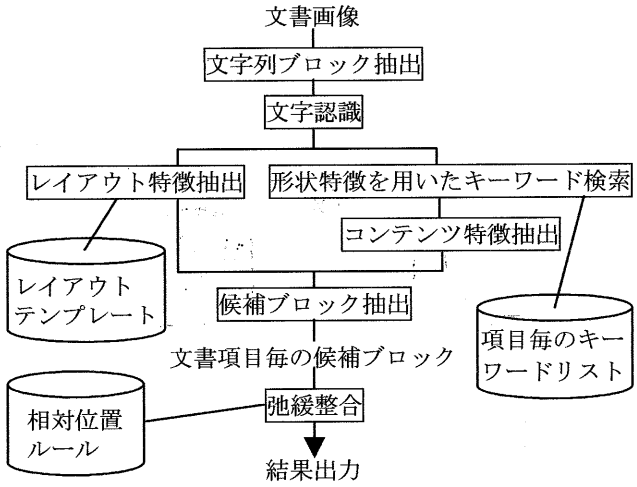


図１　処理の概要

３．レイアウト特徴

レイアウト特徴は文字列・文字切り出しの結果、得られる座標情報から算出する。レイアウト特徴は、図 2-1～図 2-3 に示す領域特徴，文字列特徴，文字特徴の３種類から構成され、各特徴ごとに平均値，分散値を特徴値として抽出する。

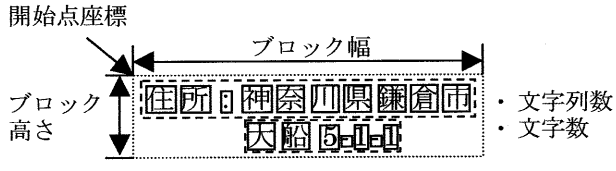


図２-１　領域特徴

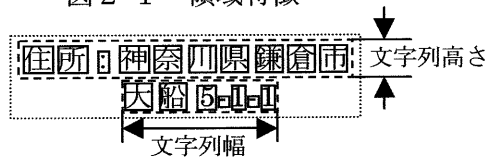


図２-２　文字列特徴

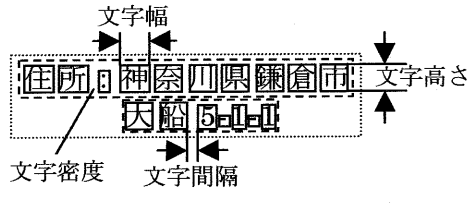


図２-３　文字特徴

A Robust Item Classification Method for Document Images
Yasuhiro Okada, Fumio Yoda
Information Technology R&D Center
Mitsubishi Electric Corporation
5-1-1, Ofuna, Kamakura, Kanagawa, 247-8501, Japan

4．コンテンツ特徴

文書項目ごとに許容されるキーワードを分類してキーワードリストに登録しておき、文字列ブロック内の文字を認識した結果得られる認識候補文字と形状特徴をもとにキーワードとの照合をとる。認識候補文字と形状特徴を併用したキーワード照合フローを図３に示す。これにより、項目毎のキーワードが文書内のどの位置に存在するのかを求め、コンテンツ特徴とする。

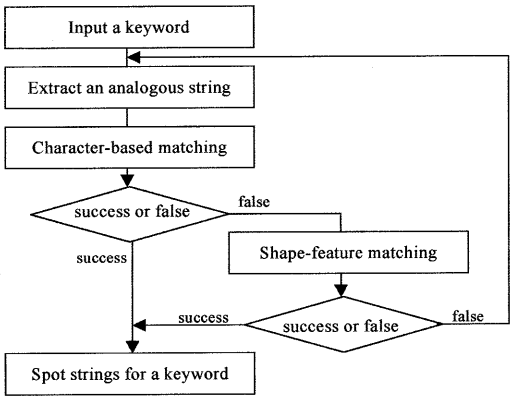
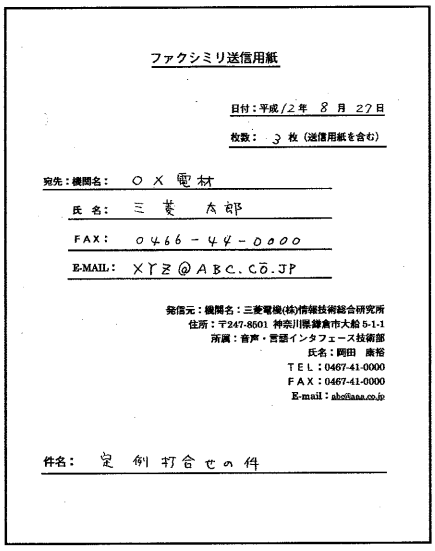


図３　形状特徴併用検索方式の処理フロー

5．実　験

5．1　評価データ

JEIDA規格[3]で規定するFAXカバーシートを対象とし、規格を満たす異なる10種類のFAXカバーシートをFAXから入力したデータを用いた。図４に評価データの一例を示す。評価データの文書項目は図５に示す構成をとり、文書項目の欠落も発生する。


ファクシミリ送信用紙

日付：平成/2年 8月 27日

枚数： 3 枚（送信用紙を含む）

宛先：機関名： ○×電材

氏名： 三菱 太郎

FAX： 0466-44-0000

E-MAIL： XYZ@ABC.CO.JP

発信元：機関名：三菱電機(株)情報技術総合研究所
住所：〒247-8501 神奈川県鎌倉市大船 5-1-1
所属：音声・言語インタフェース技術部
氏名：岡田 康裕
TEL：0467-41-0000
FAX：0467-41-0000
E-mail：abc@aaa.co.jp

件名： 定例打合せの件

図４　FAXカバーシートの例

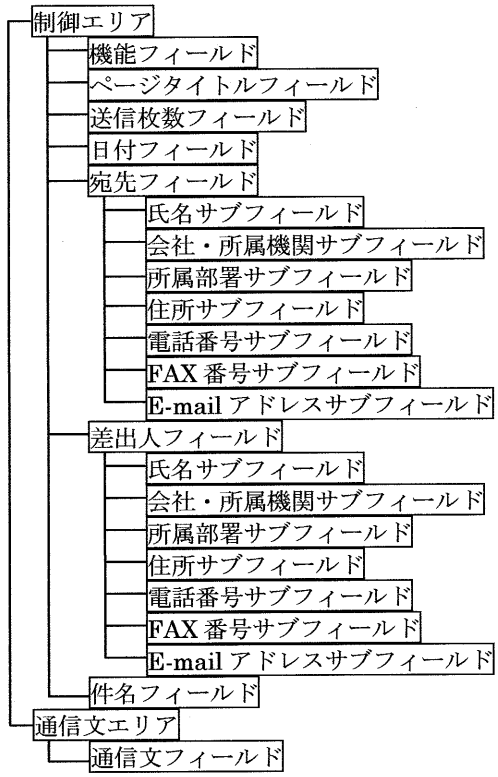


図５　FAXカバーシートの文書項目

5．2　評価結果

10種類のFAXカバーシートに対する正項目分類比率を表１に示す。全体で約94%の正項目分類率を得た。項目の湧き出しは発生しなかった。

表１　正項目分類比率

記入形式	印刷	手書き	計
正項目分類比率	115/120	24/27	139/147

6　おわりに

レイアウトと記述内容双方の情報を用いた文書項目分類方式をレイアウト変動の大きいFAXカバーページの項目自動分類に適用し、有効性を確認した。今後は、本方式を他の文書に適用する予定である。

参考文献

(1) 岡田他:"レイアウト情報と記述内容による文書項目の自動分類"信学総大 D-12-27(1997)

(2) T.kameshiro, et al. "A Document Image Retrieval Method Tolerating Recognition and Segmentation Errors of OCR using Shape-Feature and Multiple Candidates" Proc. ICDAR99 pp681-684 (1999)

(3) JEIDA規格「ファクシミリサービス高度化のためのカバーシートフォーマットの規格」,JEIDA-60-2000,(2000)