レイアウト解析による文書画像分類法の検討

3T－7

道坂 修　吉野 順　岩城 修
NTT データ通信 情報科学研究所

1 はじめに

光ファイリングシステムに代表される、オフィス文書のラスターイメージを蓄積する文書管理システムでは、文書の検索の効率化を考慮して、文書の種類に応じて自動分類する機能が不可欠である。

文書画像から文書の種類を識別する手法として、帳票と対象とした手法 [1] があるが、一般文書への適用が困難である。

そこで、分類の対象を、一般の文書の中でも特定のレイアウトを持つような準定型文書に限定し、文書のレイアウト情報とテキスト情報の双方の特徴を表現したフォーマットモデルを用いることで、分類可能な文書のバリエーションを広げることが考えられる。

本稿では、提案手法の実現方法およびその有効性について報告する。

2 フォーマットモデルの定義

文書の中にはページ中の特定の位置に固定的な文字や記号などが印字されることが多い。そこでこうした「固定記述領域」に着目したフォーマットモデル (図 1) を構築し、分類のためのフォーマット情報とする。フォーマットモデルは以下のクラスから構成され、各ページ毎に固定記述領域を一意に表現する。

Document Class 文書画像全体。属性としてページ画像と、画像サイズ (w, h) と解像度、領域数、以下に用いる領域クラスを持つ。

Region Class テキストや表／図などの領域あるいは表におけるセルの領域。属性として領域座標 (x, y, w, h) と領域属性 (Text Class ／ Table Class ／ Figure Class) を持つ。

Text Class 領域がテキストであるクラス。属性としてテキストを持つ。

Table Class 領域が表であるクラス。属性としてセル領域数と Text Class を持つ。

Figure Class 領域が図であるクラス。属性として当該領域のラスタイメージを持つ。

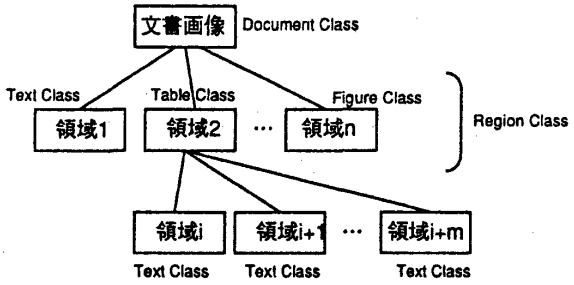
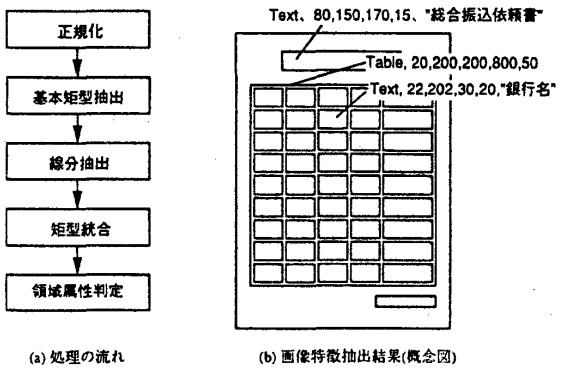


図 1: 文書のフォーマットモデル

3 フォーマットモデルの抽出および未登録文書の分類方法

同一フォーマットの文書であれば、それぞれについて「固定記述領域」が共通することを利用して、分類するフォーマット毎に以下の処理により、フォーマットモデルを得る。

1. 同一フォーマットに分類できる文書を数枚取得し、それらについて、図 2(a) に示す処理フローにより、各ページ画像の属性付き領域情報 (図 2(b)) を得る。
   (a) ノイズを除去し、傾きを補正する (正規化)。
   (b) 黒連結矩形および線分を抽出する。
   (c) 処理 (b) で得られた矩形に対し、矩形間の距離や包含関係を考慮し、矩形の統合を行う。
   (d) 統合された矩形に対し、内部の線分や黒連結矩形の規則性に着目し、テキスト／表／図の属性付けを行う。
   (e) テキスト領域については文字認識を行う。表領域については、表の罫線より表内のセルを抽出し、セル内の文字認識を行う。



(a) 処理の流れ　(b) 画像特徴抽出結果(概念図)

図 2: フォーマットモデルの要素抽出

2. 次に、各ページ画像における属性付き領域情報について、以下の処理を行う。

A method of document images classification by layout analysis
Osamu DOSAKA, Jun YOSHINO, Osamu IWAKI
NTT DATA Corporation
Email:dosaka@lit.rd.nttdata.jp

(a) 属性(テキスト／表／図)の等しい領域C_iについて式(1)の条件を満たすような矩形cを求める。式(1)にて、関数$S(C_i)$は領域C_iの面積、r_sは重なり具合を調整するしきい値である。

$$S(c) \geq S(C_i) * r_s \quad (0 < r_s < 1) \qquad (1)$$

(b) 上記を満たす矩形cが存在して、領域C_iの属性がテキストの場合、領域C_iに含まれる文字列(認識結果)の比較を行う。以下の条件をすべて満たさない場合には、矩形cの情報は破棄する。

- 各学習データにおける領域C_iに含まれる文字列s_iの文字列長がすべて等しい
- 各学習データにおけるs_iの各文字位置においてすべてに共通する文字が1つ以上存在する

(c) このようにして求めた領域cの座標情報／属性、そして領域cの属性がテキストの場合の文字列情報をフォーマットモデルのインスタンスとして登録する。

上記で得られた複数フォーマットモデルに基づき、文書の分類は以下の過程で行う。

1. フォーマットモデルの抽出と同様にして、文書画像の属性付き領域情報を得る。
2. 各領域について、登録されている各フォーマットモデルのRegion Classのインスタンスと比較し、各モデルに対する得点付けを行う。

   得点の付け方は、フォーマットモデルの各領域との重なり具合と、領域の属性がテキストの場合の文字列の一致率から算出する。

   nをRegion Classのインスタンス数、S_iをRegion Classの領域と重なる面積、SをRegion Classの領域の面積の総和として、領域C_iにおける文字列一致関数$T(C_i)$を以下の式

$$T(C_i) = \begin{cases} 1 & \text{図／表の場合} \\ \dfrac{\text{一致した文字数}}{\text{全文字数}} & \text{テキストの場合} \end{cases} \qquad (2)$$

   のように定義することにより、各モデルに対する得点tを以下のようにして求める。

$$t = \frac{\sum_{i=0}^{n} S_i T(C_i)}{S} \qquad (3)$$

3. 上記の演算にて、得点の最も高いモデルを該当する文書種別とする。

4 評価実験

以下に定義する文書識別率を求める。

$$\text{文書識別率} = \frac{\text{正しく識別された文書数}}{\text{全文書数}} \qquad (4)$$

4.1 評価環境 / 評価用データ

1. 文書のページ画像を400dpiの2値画像として入力する。
2. 評価対象文書には、オフィスで用いられる回覧文書などを5種類選択した。これらには、固定位置に共通したテキスト／表／図が記述されている。
3. 文書分類率の測定には、Leaving-one-out法[2]を用いる。1文書種別に該当する全文書10枚のうちの9枚を学習用、残り1枚を識別用として使用し、学習用／識別用の文書セットのすべての組合せについて、測定を実施した。

4.2 実験結果

Leaving-one-out法で試行を行ったすべてのケースについて、本実験では100%の識別率が得られ、良好な結果を得ることができた。

5 考察

一つの文書種別に登録される文書のセットにおいて、フォーマットが異なるものが混在していた場合、提案手法では一致するフォーマットモデルの領域数や領域面積が減少するため、誤識別を起こすことが考えられる。また、文書画像スキャン時の傾きや倍率変動により、各領域の座標値に変動が生じるケースがあり、誤識別の原因となることが想定されたが、本実験では良好な結果が得られた。

そこで今後は、学習して得たフォーマットモデルの再検証を行い、フォーマットモデルの信頼性を高めることや、フォーマットモデルの領域の座標値には、傾きや倍率によらない相対的な座標を用いることも考えられる。

6 おわりに

レイアウト情報とテキスト情報の双方の特徴を表現したフォーマットモデルを用いることにより、一般の文書の中でも特定のレイアウトを持つような準定型文書を正確に分類することができ、提案手法の有効性を明らかにした。

参考文献

[1] 徳升他: “フィールド情報に基づく帳票識別の一検討,” 信学論,PRU88-114

[2] K.Fukunaga: “Introduction to Statistical Pattern Recognition,” Academic Press, 1972