I_044

線分の境界線を用いたPDF文書からの矩形枠抽出
A Method of Recognizing Rectangles from PDF Documents

佐藤　仁†
Hitoshi Satoh

1. まえがき

近年，地方自治体や企業において，各種申請書の電子化に向けた取り組みが行われている．電子申請においては，従来の紙による申請書を文書作成ソフト等を用いて電子的に作りなおすだけでは不十分であり，適切な入力フィールドを付加し，タグづけする作業が必要となる．我々は，「電子申請システムのための文書構造化技術に関する研究」[1,2]を広島市立大学浅田研究室，株式会社ミウラ，日本アイビーエム(株)の三者で行っている．この研究は，文書作成ソフトで作成した雛型となる申請文書を解析し，質問-回答の関係を抽出することで，電子文書への入力フィールド付加，タグづけ，および，タグの管理を支援する技術を開発することを目的としている．対象とする申請文書は，各欄が罫線により分割された罫線文書であり，PDF形式[3]で与えられる．この文書の解析には，罫線で囲まれた欄（矩形枠）の抽出が必要となる．

紙に印刷された帳票をスキャンし，その画像から帳票の矩形枠を認識し，抽出することは広く行われている[4]．PDF文書(以下PDFとする)を対象とした場合，画像を入力とする場合に比べ，電子的に提供されていることの利点がある一方で，PDF固有の問題点も存在する．

本稿では，PDFから矩形枠を抽出する際の問題点と，それに対処するために考案した線分の境界線を用いた矩形枠抽出手法について述べる．

2. PDFフォーム作成支援システム

開発中のPDFフォーム[5]作成支援システムの処理手順を以下に示す．

(1) PDFを解析する．罫線を構成する線や文字の位置，各種属性を読み取り，矩形枠を認識し，その幾何的構造をXML文書として出力する．この情報を幾何情報と呼ぶ．
(2) 幾何情報に基づき，書式構造文法を用いて文書の枠同士の関係を解析する．解析結果を文書構造情報とする．さらに，文書構造情報から，入力文書がもつ質問-回答の関係を抽出する．これを指示情報と呼ぶ．
(3) 指示情報，文書構造情報，幾何情報に基づき，入力文書(PDF)に入力フィールドとそのタグ名を付加する．同時に，文書中に表れる質問-回答の関係と入力フィールド，タグ名との対応関係の情報を出力する．

本稿は，上記の(1)の手順中の矩形枠の抽出を扱う．

3. PDFの矩形枠の構成要素

矩形枠を抽出するためには，矩形枠がPDF文書内でどのような形式で表現されているかを知る必要がある．いくつかのPDF文書を解析した結果，矩形枠には3種類あることが分かった．

以下の説明中のRECT要素及びLINES要素はPDFを記述するための要素であり，RECT要素が矩形を表現し，LINES要素が複数の線の集合を表現する．

I. RECT要素の境界線が矩形枠を表現しているもの
II. 多数の線分が集まって矩形枠を表現しているもの
III. RECT要素の境界線と線分とが組み合わされて矩形枠を表現しているもの

我々の扱った実際のPDFでは，IIの型が多く見られた．また上記説明中の線分についてはさらに次の二種類があることがわかった．

① 細長いRECT要素の内部を塗りつぶしたもの
② LINES要素がコの字形に連結された3本の線の集合となっており，その始点と終点を結んで得られる細長い長方形の内部を塗りつぶしたもの

図1は，PDFに罫線がどのような形で入っているかを誇張して示した例で，PDFの各要素を表示したものである．この例などから次のこともわかった．

(1) 1本の線分に見えても複数の線分が連結されている（図1のD1とD2）
(2) 目視ではわからないわずかの隙間がある場合がある（図1のD1とD2，あるいはD2とE）
(3) 1本の線分に見えても太さがわずかに異なった線分が連結されていることがある（図1のD1とD2）
(4) RECTからなる線分とLINESからなる線分が連結されている場合がある（図1で例えばD1がRECT，D2がLINESの場合がある）

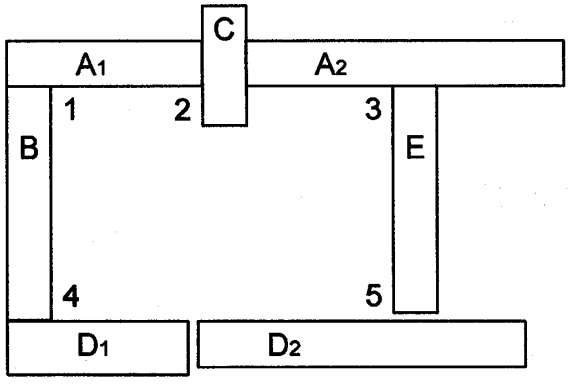

図1PDFの罫線の例を誇張したもの

4. 矩形枠抽出での画像とPDFとの違い

矩形枠抽出での，画像とPDFとの違いを次に述べる．

†日本アイビーエム(株)，ソフトウェア開発研究所

(1) 画像の場合は全体がわずかに斜めになっている場合があり，スキュー補正が必要である．PDF では線分は水平または垂直であり補正は不要である．
(2) 画像の場合は読み取り部の条件の違いなどから線の太さなども正確に同じではない．PDF の場合にも1本の線分が異なった太さの線分の連結からなる場合があるが，計算誤差由来と思われる程度の差である．図 1 の D1,D2 はその誇張した例である．実際にはその差は目視できない．
(3) 画像の場合やはり読み取り部の条件の違いなどから線がかすれ，1本の線が分離してしまう場合がある．PDF の場合には二通りの分離のしかたがあり，1) 計算誤差由来と思われるわずかの隙間の場合，2) 文書を作成する際の作図の問題や PDF 変換プログラムが分離して描画することからくる，目視できる程度の分離の場合がある．図 1 の D1,D2 はわずかに分離した例であり，その差は実際には目視できない．また，A1,A2 は線分 C を挟んで目視できる程度に分離している例である．

次に PDF に固有の問題点をあげる．
(4) 細い線分の上を太い線分が覆い隠す場合があり，覆い隠されて見えない線分が矩形枠の抽出に関わってはいけない．画像についてはこのような問題は生じない．
(5) 背景と同色か不可視属性を持つ線分は，矩形枠の抽出に関わってはいけない．画像では最初から認識されない．
(6) PDF の場合先述の通り 3 種類の矩形枠(Ⅰ,Ⅱ,Ⅲ)があり，それぞれの処理を行わなければならない．画像の場合には線分のみからなるとして処理する．

5. 矩形枠抽出

5.1 矩形枠抽出の要求仕様

矩形枠を抽出するにあたっての主な要求仕様は以下のようなものであった．
(1) 表などの場合，内側の矩形枠だけを抽出し，それらが複数組み合わされた矩形（例えば表全体の外枠）は抽出しないこと
(2) 入れ子になっている矩形は XML 文書作成の際，内側の矩形要素を外側の矩形要素の子要素とすること
(3) 矩形は位置とサイズを持つが，その座標については線分の中心線を用いること

5.2 矩形枠抽出に境界線を用いた理由

矩形枠の抽出は線分の境界線(左右天底の4種類)を用いた．矩形枠抽出の要求仕様では矩形の座標に線分の中心線を用いることになっていたにもかかわらず，矩形枠抽出に境界線を用いた理由を述べる．

最も大きな理由は，4.の PDF 固有の問題点の(4)である．境界線を用いた場合には，覆い隠された線分は抽出された矩形枠には含まれない．覆い隠している線分による矩形枠の方が先に見つかるからである．しかし，中心線を用いた場合には，そうならない場合がある．図 2 で中心線により抽出を行った場合には線分 BCDE による矩形は，線分 ACDE による矩形の内側となり抽出される．しかし，線分 A は線分 B を覆い隠すので線分 ACDE による矩形が抽出されるべきである．境界線を用いた場合にはそのようになる．

その他に長軸と短軸の長さが同程度の正方形に近い線分の場合どちら向きの線分とすべきか難しいが，境界線による場合はそのような問題は生じない．

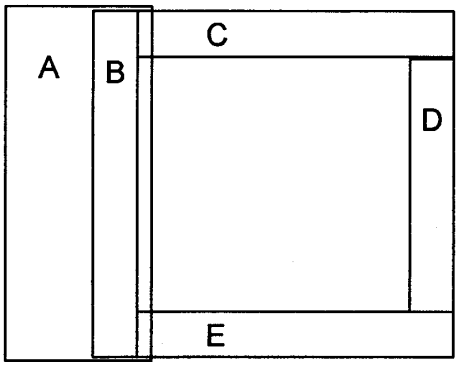

図 2 細い線の上を太い線が覆い隠す例

5.3 矩形枠抽出の詳細手順

PDF についての調査，及び要求仕様を踏まえ，矩形枠の抽出手順を次のように実装した．ただし記述を簡単にするために，Ⅲの型の矩形枠の存在は無視している．
(1) 関連する PDF 要素(RECT 要素と LINES 要素)を読み取る．不可視属性を持つもの等は除く．
(2) RECT 要素についてさらに分類する．
　① RECT 要素が境界線のみで，内部を塗りつぶさないものはそれ自身を独立の矩形枠として登録する．
　② 内部を塗りつぶす RECT 要素は線分として登録する．
(3) LINES 要素については内側を塗りつぶす場合にそれを線分として登録する．
(4) 可能な限り線分を連結する．ここで連結とは複数の線分の代わりに，合成した一つの線分を登録することを言う．数ポイント程度の隙間があっても，線の太さや座標にわずかのずれがあっても連結する．線分は RECT 要素からなるものと LINES 要素からなるものとを区別しない．ただし色が異なる場合は連結しない．以下線分とは連結された線分のことである．図 1 で A1,A2 は連結され線分 A となり，D1,D2 は連結され線分 D となる．
(5) 線分どうしの交点の4種類の隅(コーナー)についての情報を登録する．全てが存在するとは限らない．二つの線分 X と線分 Y について考える．
　① 線分 X の天と線分 Y の左からなる右下隅
　② 線分 X の天と線分 Y の右からなる左下隅
　③ 線分 X の底と線分 Y の左からなる右上隅
　④ 線分 X の底と線分 Y の右からなる左上隅
　また接触するか否かの判定は少し余裕をもたせなければならない．例えば接触していない場合には交点は存在しないが，数ポイント程度の隙間ならば接触しているとみなして交点を求める．図 1 で線分 D と線分 E の間に隙間があるが，接触しているとみなす．
(6) 矩形の探索を行う．同一の線分を構成要素に持つ異なった種類の隅は，その共通の線分により繋がっている．
　① **左上隅に繋がる右上隅を見つける．**

② **左上隅に繋がる左下隅を見つける.**
③ **右上隅と左下隅に繋がる右下隅が見つかればそれが抽出された矩形となる.**

図1の例では，左上隅1は右上隅2と3に繋がる．同時に左上隅は左下隅4に繋がる．そこで右上隅2と左下隅4に繋がる右下隅はないが，右上隅3と左下隅4に繋がる右下隅5が見つかり，線分ABDEによる矩形が抽出される．

上記手順は内側から順に探索するので，最初に見つかった矩形が最内部の矩形となる．

5.4 結果

上記実装はPDFフォーム作成支援システムに組み込まれ，良好な矩形枠抽出結果を得ている．図3に本支援システムにより作成されたPDFフォームの例を示す．

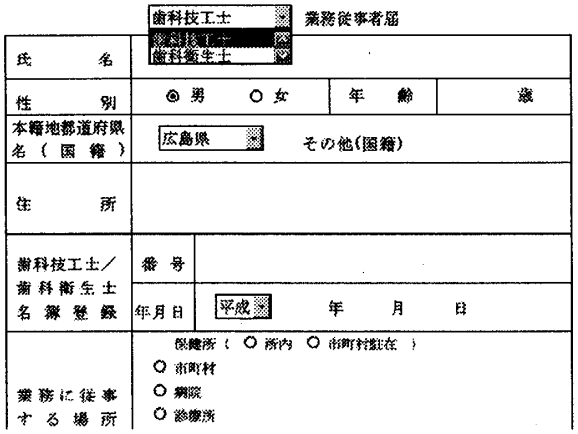

図3 作成されたPDFフォームの例

矩形枠抽出のテストは84種類の実際の申請書形式のPDFについて行い，矩形枠の抽出に失敗したものは1件であった．失敗した1件は異なった太さの線分が連結されなかった結果であり，線分連結の際の許容する太さの差を拡大することで解決するものであった．

6. まとめ

本手法はPDFフォーム作成支援システムに用いられ，良好な矩形枠抽出結果を得た．境界線を用いて矩形枠を抽出する手法は，PDFにおいて矩形枠を抽出する場合のPDF固有の問題を解決する．

謝辞 本稿を作成するにあたり貴重な御助言を頂いた広島市立大学浅田尚紀教授，同椋木雅之助教授，同青山正人助手，(株)ミウラ浅木森友彦氏，日本アイビーエム東京基礎研究所杉本和敏氏に感謝する．本研究は総務省の戦略的情報通信研究開発推進制度の支援を受けたものである．

参考文献

1) 浅田尚紀, 椋木雅之, 青山正人, 浅木森友彦: 電子申請システムのための文書構造化記述の研究. PRMU2005-1111, 2005-11.
2) 青山正人, 小柴和宏, 椋木雅之, 浅田尚紀: 書式構造文法を用いた表構造を含む対話型罫線文書の解析, 画像の認識理解シンポジウム MIRU2004, vol.I, pp.I-327-332, 2004-07.
3) http://www.adobe.co.jp/products/acrobat/adobepdf.html
4) Y.Y Tang et al.: Multiresolution Analysis in Extraction of Reference Lines from Documents with Gray Level Background. IEEE PAMI 19, 8, pp921-925, 1997.
5) http://www.adobe.co.jp/epaper/eform/main.html