I-001

寸法不変な特徴量を用いた帳票レイアウト解析

Form layout analysis using size invariant features

鈴木 智久†
Tomohisa Suzuki

あらまし 帳票画像の文字認識においては、枠検出により認識対象領域の位置を特定する必要がある。その際には紙面上の固定座標を用いる場合が多い。しかし、給与支払報告書など一部の一般帳票については枠の寸法が厳密に定義されていないため、固定座標による枠検出は困難である。本稿では、枠の構造的な特徴量に基づいたテンプレートマッチングを行い、枠を検出する手法を提案する。本手法により、寸法の変動に対して頑健な枠検出が可能である。実験では給与支払報告書２００枚の９０％から正しく枠を検出できた。

１．まえがき

帳票画像上の文字を認識する場合、通常、枠を検出し、認識対象領域の位置を特定する必要がある。ＯＣＲ帳票のように枠の寸法が一定な帳票の場合は、紙面上の固定座標を枠位置とする方法が一般的である。一方、給与支払報告書に代表される、枠の寸法やデザインが多種に渡る一般帳票の場合には、枠位置として固定座標を用いるのは不適切であり、レイアウト解析により枠を検出する必要がある。給与支払報告書の例を**図１**に示す。全国の自治体が多数の企業・法人から収集することから、その枠の寸法、デザインは多種多様である。これまで、このような一般帳票のレイアウト解析技術としては、テンプレートとのイメージマッチングに基づく方法、罫線線密度による位置あわせを行う方法等が提案されているが、これらの方法では枠線の寸法の変動に十分には対応できない。また、枠の見出しを認識し枠位置を特定する方法も提案されているが[1]、かすれや文字接触による見出し文字の誤認識に影響されることがある。

本稿では、同種の一般帳票においては、デザインにより記入枠の寸法が違っていても記入枠同士の位置関係がほぼ一定であるという事実に注目し、寸法変動に対して頑健なレイアウト解析手法を提案する。記入枠の寸法とは無関係な構造的な特徴量を抽出し、その特徴量の評価によりレイアウト解析を行うという考え方である。具体的には、レイアウト解析の対象となる領域候補とテンプレートの該当箇所のそれぞれの特徴量の残差二乗和が最小となるよう、ＤＰマッチングを用いて領域分割する。寸法とは無関係な特徴量と非線形なマッチングにより、寸法の変動に頑健なレイアウト解析が期待できる。

図1：寸法の異なる様式の例
受給者生年月日の欄の寸法の違いが顕著である。

2. 提案手法

2.1.レイアウト解析の概要

提案手法では、帳票の様式を**図２**に示すように縦方向に分解した「段」と、一部の段を横方向に分解した「カラム」という単位を考え、それらの位置あわせを行うことにより記入枠の位置の特定を行う。

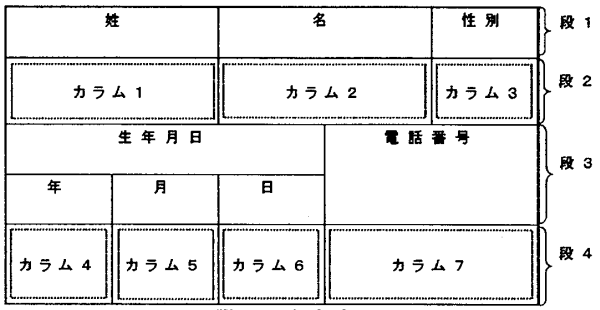


図２：段とカラム
様式を縦に分割した単位を「段」、一部の段を横方向に分割した単位を「カラム」と呼ぶ。段への分割の様子を右側に、カラムを点線で示す。

提案手法による枠位置の特定は以下の流れで行う：

1. 段の位置合わせ
2. カラムの位置合わせ
3. カラムからの記入枠の検出

段、カラムを逐次的に特定しているのは、各処理が後続の処理に依存せず、独立に実行できると期待できるからである。

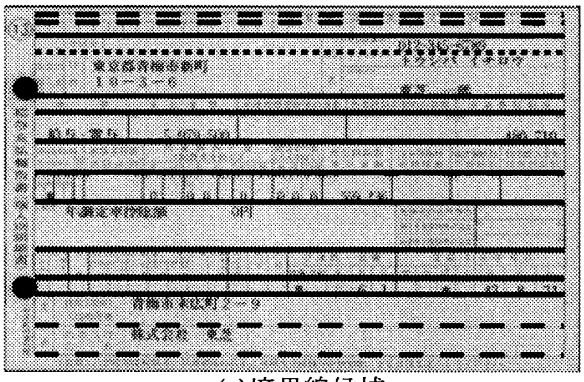
以後、**2.2節**では段とカラムの候補とそれによるマッチ

† 東芝ソリューション株式会社

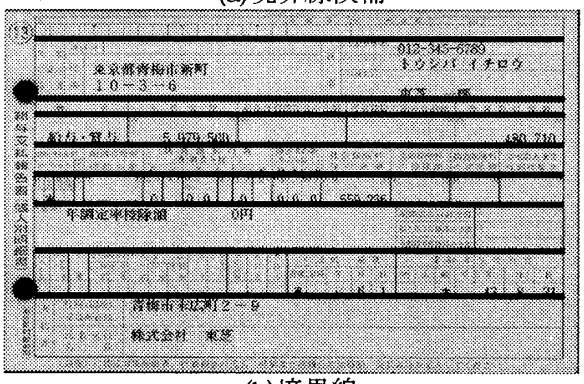
ング処理の必要性について、**2.3 節**では寸法に依存しない特徴量とそれによる領域候補選択の評価について、**2.4 節**ではＤＰマッチングによる領域候補選択方法、**2.5 節**ではカラムからの枠検出の方法について述べる。

2.2.段とカラムの候補

段の境界は通常長い横罫線で区切られている為、**図２**のような単純な様式の場合には、長い横罫線の検出のみで段の位置あわせが実現できる。しかし、給与支払報告書では、処理対象範囲の枠配置が定まっている反面、処理対象の上下に配置された枠がデザインにより異なる為、それらの枠の横罫線は位置あわせに用いない方が良い。また、枠配置が定まっている部分においても段の境界とならない横罫線が検出される為、横罫線の検出のみでは段の境界を特定できない。従って、段位置を推定するには、横罫線の検出により検出した境界候補の系列と、テンプレート上の境界をマッチングし、境界候補を選別する必要がある。



(a)境界線候補



(b)境界線

図3:境界線候補と境界線

処理対象範囲外の枠線（(a)破線）や、処理範囲内でも検出する段の境界とならない罫線（(a)点線）も余分な境界線候補として検出される。

また、検出した段からのカラムの推定においては、縦罫線をカラムの境界候補とすることが可能であるが、**図４**の例のように縦罫線の検出失敗等が起きることがあり、縦罫線の系列はカラム境界の系列と完全には一致しない。従って、カラムの位置合わせにおいても境界と境界候補のマッチングが必要となる。

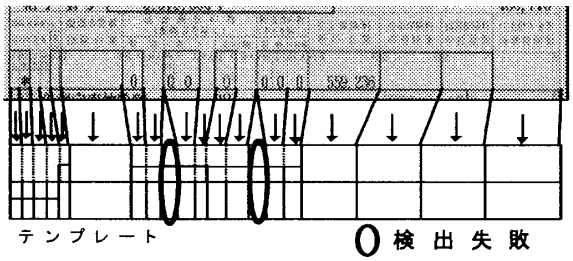


図4:カラム境界の検出失敗例

カラムの境界線が点線であったり省略されていたりすると検出に失敗することがある。

2.3.特徴量による段･カラムの評価

以下では、テンプレート上の境界の間の領域を「領域」、帳票画像上の境界候補の間の領域を「領域候補」と呼ぶことにする。そして、領域および領域候補の各々の特徴量の類似性によってそれらの一致の程度を評価する。

段の特徴量としては段の上端に接する縦罫線の本数と、段の下端に接する縦罫線の本数(**図５**参照)を用いる。この特徴量は寸法変動の影響を受けない。

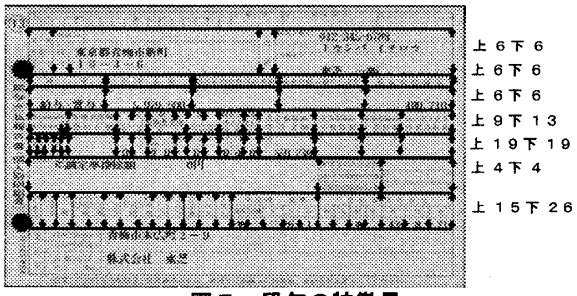


図５：段毎の特徴量

段の上下の端に接する縦罫線の本数を右側に示した。矢印で縦罫線の接触箇所を示した。寸法に依存しない特徴量が抽出されている。

また、カラムの特徴量としてはカラムの左右端を上に伸ばした線とカラム直上段の上下端からなる矩形の四隅に接する縦罫線の本数（**図６**参照）を用いる。これらは寸法によらない特徴量である。カラム内ではなくカラム直上の領域の特徴を用いたのは、カラム内のデザイン、特にプレプリント文字の配置がまちまちで、カラム内の画像からは安定した特徴量が求められないからである。

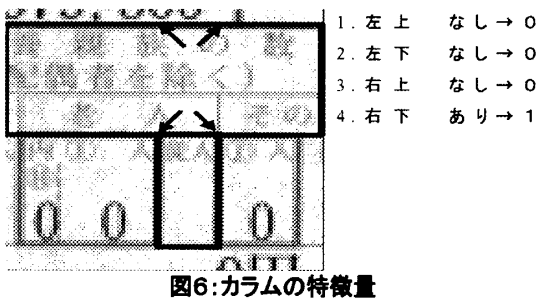


図6:カラムの特徴量

カラム直上の段の縦罫線の有無を注目カラムの左右の端について上下二通り調べ、その有無により0－1の特徴量を割り当てる。

2.4.DP マッチングによる段･カラムの推定

段・カラムいずれの位置あわせもテンプレート上の境界の１次元系列と、境界候補の１次元系列について行われる為、位置あわせにはDPマッチングを用いる。

図７は段の位置あわせでの DP マッチングの一例である。経路上の格子点が検出された境界線候補と段の境界の対応付けを表し、格子点同士を結んでいる辺が領域と領域候補の対応付けを表している。また、左下、右上の水平部分は上下数本の枠線の無視を意味している。

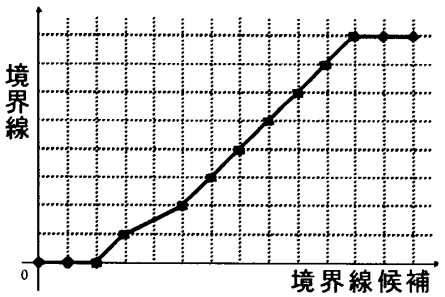


図7:段の境界線と境界候補のマッチング

DP マッチングにおいては、各辺についてコストを定義する必要があり、提案手法では各辺での領域と領域候補の特徴ベクトルの距離の二乗和をコストEとした。領域の個数をN、i番目の辺に対応する領域の特徴ベクトルを$\mathbf{x}_i$、領域候補の特徴ベクトルを$\mathbf{y}_i$とすると、コストEは以下

$$E=\sum_{i=1}^{N}\left\|\mathbf{x}_i-\mathbf{y}_i\right\|^2$$

の式で表される：
このコストが最小になるように領域候補が選択される。

2.5.カラムからの記入枠の検出

カラムからの記入枠は、カラムの左右の端に達している横罫線の内、カラム下端から２本を検出し、それらを上下端とする矩形として検出した。（**図８**参照）

この方法では検出できない例外的な枠については、それらに固有な処理を実装したが、それらの詳細については割愛する。

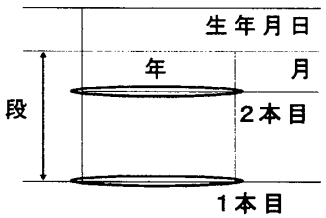


図8:カラムからの記入枠の検出

カラムの左右の端に達している横罫線のうち下から二本を記入枠の上下端とする。

3. 実験結果

図 9 は提案手法における段の位置あわせの成功例である。図の矢印は段の境界を示している。**図１０**はカラムの推定結果の一例を示している。

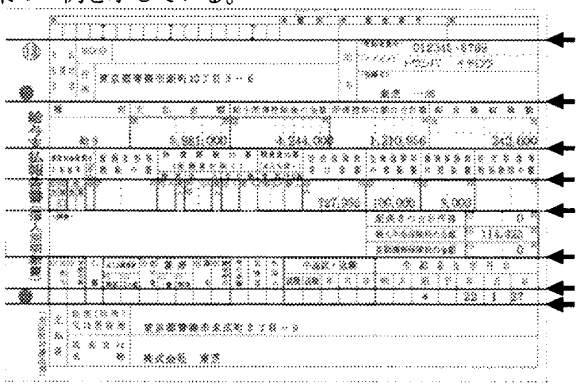


図 9:段の検出結果

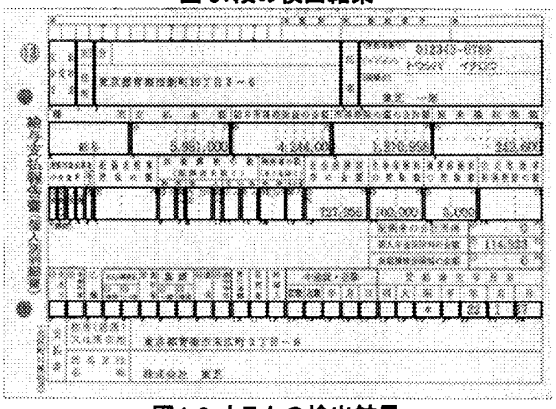


図１０:カラムの検出結果

無作為に選出した給与支払報告書２００枚のスキャナ入力画像について提案手法での枠検出を行ったところ、１種類のテンプレートで９０％の帳票からの枠検出に成功した。

4. まとめ

本稿では、罫線枠で構成された様式に有効な枠位置推定方法として、特徴ベースのテンプレートとのマッチングによる帳票レイアウト解析手法を提案した。複数デザインが混在した給与支払報告書への適用実験において１種類のテンプレートで９０％の帳票からの枠検出に成功したことにより、提案手法が様式変動に対し頑健であることが確認できた。

今後は性能向上の為、マルチテンプレート化を試みるとともに、見出し文字列の認識など従来技術の取り込みを行う予定である。また、適用範囲を拡大するため、他の帳票様式にも対応する予定である。

参考文献

[1]宇田：表形式既存帳票認識システム．FIT2002 pp.167-168