

検証処理に基づいた帳票の構造認識法とその評価*

6K-8

祖父江恒夫† 渡邊豊英†

名古屋大学大学院 工学研究科 情報工学専攻‡

1 はじめに

文書画像から有意な情報を自動的に抽出することを目的とした課題に文書画像理解があり、今日までに種々様々な文書に対して研究されてきた。これらの多くは、文書に関する知識を予め文書モデルとしてシステムに保有させ、その文書モデルに基づいて文書画像を解釈する枠組みを採用している。帳票文書を対象とした研究も、その応用域の広さから数多く報告され、その一つに成瀬らの試みがある[1]。成瀬らの研究では、帳票の項目域は実線の枠罫線から構成されるとして、構造記述木で表現された文書モデルの下に構造認識法を提案している。しかし、実際には各々の項目域が完全に実線の罫線で囲まれていない帳票も多く存在する。罫線が実線ではなく、破線であったり、罫線がない場合がある。このような帳票を成瀬らの方法に従って処理するためには、罫線を補足する必要がある。もちろん、その補足処理は正しく実行されなければならない。

本稿では、種々様々な帳票に対して成瀬らの方法を適用可能にする罫線の補足処理について報告する。我々の補足処理に対する視点で、特徴的なアプローチは補足した罫線に対して検証を行うことである。また、この補足処理によって拡張された方法について、その有効性を実験結果にて明らかにする。

2 帳票の構造認識

帳票の構造認識は図1に示すように処理される。ここで、罫線の補足処理は構造解析処理の一部として組み込まれ（構造解析の前処理）、補足された罫線は構造解析処理の検証手続きにより、認識上の正当性が調べられる。

構造解析処理では、ボトムアップに帳票を処理し、縦・横方向の罫線で囲まれた項目域の左上角を抽出する。すなわち、このように抽出された左上角は項目域に一対一に対応付けられる。一方、構造解析処理では構造

記述木と呼ぶレイアウト知識を用いて、トップダウンに左上角相互の構造的な関係を把握する。

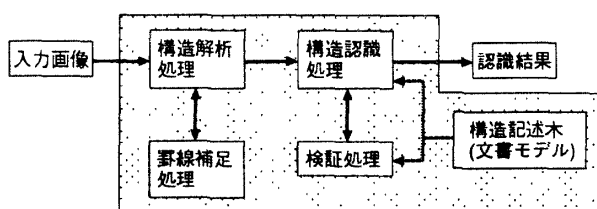


図1: 処理の流れ

3 罫線の補足処理

本構造認識では項目域の左上角が構造解析処理で重要な役割を有しており、罫線で囲まれていない、または実線でない罫線で囲まれている項目域に対して、実線の罫線を付加し、構造解析処理で項目域に対応する左上角を抽出できるように原帳票を変換するのが、罫線の補足処理である。例えば、図2に示すような処理である。罫線を帳票のどこに補うかは帳票構造が正確に理解されていないと困難であるが、ここでは帳票構造の概略特徴（ヒストグラムなどから）に基づいて、その位置などを推定し、仮に設定するという前提で処理する。従って、それが実際に正しいか否かは、構造解析処理でレイアウト知識の下に検証される。

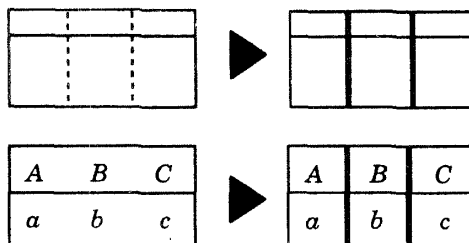


図2: 罫線の補足処理

4 補足処理の検証処理

罫線の補足処理で仮に設定された罫線が、帳票構造を解釈するときレイアウト知識と整合するか否かを調べ

*Structure Recognition Method Based on Validation of Table-form Documents and Its Evaluation

†Tsuneo SOBUE and Toyohide WATANABE

‡Department of Information Engineering, Graduate School of Engineering, Nagoya University

必要がある（もちろん、この場合、レイアウト知識は原帳票そのものの構造知識ではなく、認識すべき項目域から構成された帳票のレイアウト知識となっている）。

検証処理では、項目域間の従属関係に基づいて、認識結果を検証する。ここで、従属関係とは意味的に関連する項目域間の関係のことであり、例えば図3(矢印が従属関係を表す)に示すように項目データ域は項目名称域に従属する。従属関係にある項目域群は必ず矩形域になる。よって、構造解釈処理で解釈された結果において、従属関係にある項目域群が矩形域になっているか否かを確かめ、もし矩形域になっていない項目域群がある場合には、その項目域群を矩形域に修正し、再び構造解釈処理を実行する。項目域間の従属関係がすべて正しく認識されるまで、検証・再解釈を繰り返す。例えば、図4では、構造解釈処理の結果に対して、従属関係にある項目域群(項目域Aと項目域a)が矩形域になっていないので、これらの項目域群を矩形域に修正し、再解釈する。ただし、この再解釈では、前の検証処理で項目域Aと項目域aは従属関係により正しく確定されているので、これらの項目域以外の項目域について解釈する。同様に、この再解釈された結果に対して、従属関係にある項目域群(項目域Bと項目域b)が矩形域になっていないので、それらの項目域群を矩形域に修正し、再解釈する。この再解釈された結果は、従属関係の項目域群がすべて矩形域になっているので、これを認識結果とする。このように検証・再解釈を繰り返すことで、罫線を補足したことによって誤って項目域が確定されてしまった構造を正しく修正することができる。

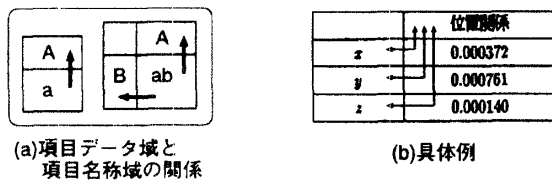


図3: 従属関係

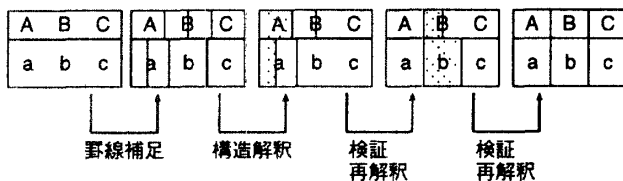


図4: 検証処理の流れ

5 実験結果

本稿で述べた方法にしたがって、帳票40枚に対して実験し、図5にその一例を示す。(a)は入力画像、(b)は罫線を補足した画像、(c)は検証処理をしなかった場合の画像、(d)は検証処理をした場合の画像である。そして、検証処理なしの場合と検証処理ありの場合の認識結果を表1に示す。表1において、正しい数は正しく確定した項目域数を、誤った数は誤って確定した項目域数を表している。実験結果より、検証処理をすることで、誤って確定する項目域の数を減らし、認識率を向上させることができた。

Target	Context	CR Rate	Rating
Face	Studied	.89	4.38
	No	.91	5.00
	New	.89	4.50
Voice	Studied	.53	.02
	No	.39	-1.30
	New	.66	1.17

(a) 入力画像

Target	Context	CR Rate	Rating
Face	Studied	.89	4.38
	No	.91	5.00
	New	.89	4.50
Voice	Studied	.53	.02
	No	.39	-1.30
	New	.66	1.17

(b) 罫線補足画像

Target	Context	CR Rate	Rating
Face	Studied	.89	4.38
	No	.91	5.00
	New	.89	4.50
Voice	Studied	.53	.02
	No	.39	-1.30
	New	.66	1.17

(c) 検証なし

Target	Context	CR Rate	Rating
Face	Studied	.89	4.38
	No	.91	5.00
	New	.89	4.50
Voice	Studied	.53	.02
	No	.39	-1.30
	New	.66	1.17

(d) 検証あり

図5: 実験例

表1: 実験結果

	項目域数 (A)	正しい数 (B)	誤った数	認識率 [%] (B/A)
検証なし	632	582	50	92.1%
検証あり	632	624	8	98.7%

6 おわりに

本稿では、検証処理に基づいた帳票の構造認識法を説明し、実験結果による評価について報告した。実験を通して、この手法が種々様々な種類の帳票にも対処でき、高い認識率を達成できることを明らかにした。

参考文献

[1] 成瀬博之, 渡邊豊英, 駱琴, 杉江昇: “枠罫線情報を用いた帳票文書の構造認識”, 信学論 D-II, Vol. J75-D-II, No. 8, pp. 1372-1385 (1992).