



Title	文書画像の構造解析とSGML文書生成システムの構築に関する研究
Author(s)	山岡, 正輝
Citation	大阪大学, 2000, 博士論文
Version Type	
URL	https://hdl.handle.net/11094/42166
rights	
Note	著者からインターネット公開の許諾が得られていないため、論文の要旨のみを公開しています。全文のご利用をご希望の場合は、 ＜a href="https://www.library.osaka-u.ac.jp/thesis/#closed">https://www.library.osaka-u.ac.jp/thesis/#closed >大阪大学の博士論文についてをご参照ください。

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

氏 名	やま おか まさ き 山 岡 正 輝
博士の専攻分野の名称	博 士 (工 学)
学 位 記 番 号	第 1 5 5 2 7 号
学 位 授 与 年 月 日	平成12年 3 月 24 日
学 位 授 与 の 要 件	学位規則第4条第1項該当 基礎工学研究科情報数理系専攻
学 位 論 文 名	文書画像の構造解析と SGML 文書生成システムの構築に関する研究
論 文 審 査 委 員	(主査) 教 授 北 橋 忠 宏 (副査) 教 授 橋 本 昭 洋 教 授 田 村 進 一

論 文 内 容 の 要 旨

紙を媒体とする文書をデータベース化するには、例えば、学术论文のタイトルや著者名など、検索の際にキーとなる論理オブジェクトを抽出する必要がある。また、著者と対応する所属機関名など、論理オブジェクト間の論理的な関係である論理構造もデータベースに格納する必要がある。このように、紙媒体の文書を光学的に読み取った文書画像から論理構造を抽出する技術を構造解析技術と呼ぶ。

従来、文書を構成する文字や図などの要素の配置関係を示すレイアウト構造を解析することにより論理構造を抽出する手法が提案されている。これらの手法は、学术论文のタイトルページや特許願などを対象としており、レイアウト構造および論理構造がページ単位で定義可能であることに着目し、あらかじめ作成したページ単位のモデルを他のページに適用することで解析を進めるものである。しかし、解析に構成要素の配置関係を用いているために、例えば図や表など位置の固定しない構成要素がある場合など、構成要素の相対位置が変動する文書は解析の対象外とされていた。特に、広告やカタログのような文書では、文書ごとに論理構造も異なるため、事前にモデルを用意するというアプローチをとることができなかった。

本論文では、まず、従来のような構成要素の配置関係に基づいた解析ではなく、画像特徴を用いた統計的解析により論理オブジェクトを抽出する手法を提案する。これは、構成要素ごとに文字の大きさや文字間隔、行間隔などが異なることに着目し、これらを画像特徴としてとらえ構成要素を識別する手法である。次に、1ページ中のある領域に対する部分構造モデルを作成し、これをくり返し適用することで文書構造の解析を進める手法を提案する。複数の製品が掲載されている広告画像を解析対象として、ひとつの製品を単位とする部分構造モデルを用いて解析を試みる。さらに、提案手法を用いて、学术论文誌から国際標準である SGML (Standard Generalized Markup Language) で記述された文書を生成するシステムを構築し、その実用性を評価する。

論 文 審 査 の 結 果 の 要 旨

電子化文書は印刷文書よりも検索や蓄積に有用であるため、既存の文書についてもその電子化を図ることに強い社会的要請が存在する。既存文書の電子化に当たり、文字認識を用いた本文あるいはデータの符号化は当然ながら、

標題・作成者などの書誌情報は文書の検索に極めて重要であるため、その抽出は不可欠である。ところで、同種の文書中では、書誌項目はそれぞれ特定の配置をとることが多い。本論文では、このことに着目し、詳細なレイアウト構造の解析結果を利用することにより、書誌項目を抽出する実用性に富む従来の文書画像処理にない2つの方法を提案している。

第一は、統計的処理を用いる手法である。処理の対象とする文書群から採取したサンプル文書についてレイアウト構造を詳細に解析し、その結果を各書誌項目ごとに分類する。その各分類について、レイアウト要素のデータを主成分分析することによって、それぞれの書誌項目を特徴づける主要なレイアウト要素のデータを決定する手法である。

第二は、対話的処理に基づく手法である。処理の対象とする文書群のサンプル文書の画像に対し、利用者が求める情報を含む領域を画面上で指定する。システムは、その領域に限定して、文書構成要素を抽出する画像処理を実行し、抽出された要素を利用者に提示する。この提示された構成要素に対して、利用者がそれぞれの書誌項目名を指定する。システムはこの指定とそれまでに得た処理結果とを関連づけることにより、予定した文書群における各書誌項目の抽出規則を修得するという手法である。

本論文の研究成果は、文書画像処理に関して従来には見受けられなかった発想に基づくものであり、既存文書の電子化のための実用性に富んだシステム開発に貴重な貢献をしている。よって応用ソフトウェア分野の発展に資するところが大きく、博士（工学）の学位論文に値するものと認める。