博 士（情報科学） 猪 村 元

学 位 論 文 題 名

A Full-Text Search System for Document Images Based on Character Shape Features

（文字の形状特徴量を利用した文書画像の全文検索技術）

学位論文内容の要旨

近年，世界各国の大学や企業，機関による大規模な図書館や公文書館の文書の電子化プロジェクトが盛んに進められ，ウェブを通じて利用可能な電子図書館サービスが重要視されている．これにより知的財産権の消滅した文書に関しては，その全文がウェブ上に公開され自由に閲覧できるようになってきた．また，これまでは閲覧が制限されてきた歴史的に貴重な文献や資料の画像としての公開も広く行われるようになってきた．

膨大な量の文書の中から目的の文書を探すには検索機能が不可欠である．しかし，文書の内容すべてにアクセスできる全文検索機能の提供は，文字認識 (OCR) 適用可能な現代の標準的な字体で印刷されたものに限られている．活版印刷時代の文書や，手書きの古文書などの歴史的文書に対する検索は，原本の経年劣化や，多様な字体の問題が存在しており，未だ不十分である．特に古文書画像においては，崩し文字や続け文字，経年劣化の影響によって正確な文字単位での切り出しが非常に困難である．これらの文書に対して，全文検索の手法を提供することは意義が大きい．

また，ウェブを介しての文書画像の利用には，ネットワーク負荷の低減技術も重要である．特に，昨今のモバイル・ネットワーク環境と端末の発展にともない，その重要性はさらに増している．

本研究では，印刷，および手書きの文書画像を対象にして，文字の図形としての形状特徴に基づいた擬似的な文字コードを文書の内部表現として用いることで高速な全文検索を実現するトランスメディア技術に基づきこれらの問題の解決を図った．これらの手法は対象とする文書画像のみから特徴量を抽出し，対象文書内で相対的に類似した形状の文字列を検索するため，特定のフォントや言語に依存せずに統一的な手法で適用可能である．

本研究における成果は以下の 3 点である:

1. 文書画像に適した画像圧縮手法と，圧縮文書に対する検索技術の提案
2. 印刷文書画像に対する M-tree 索引構造を用いた高速な文書画像検索技術の提案
3. 図形特徴に基づく手書き古文書画像の全文検索技術，および HMM 学習を用いた適合性フィードバックによる検索精度の向上手法の提案

本論文の構成は以下の通りである．第 2 章では，文書画像検索技術，文書画像圧縮技術に関する既存の研究について解説すると共に，それらにおける問題点を提起している．第 3 章では，本研究の基盤となるトランスメディア技術について解説している．第 4 章から第 6 章にて，上記の各成果について詳細を説明している．

第 4 章では，第 3 章で述べたトランスメディアの検索技術を利用した，文書画像に特化した画像圧

縮手法について提案している. 従来より, 文書画像の形式として JPEG や GIF 等の一般的なフォーマットを利用すると, 十分な品質を維持するためには大きな保存容量が必要となることが問題視されてきた. 本研究では, トランスメディアの検索技術を利用することで, 同一文書画像中における同じ文字の出現を見つけることが可能なことを利用して, これらを 1 つの代表画像と代表擬似コード, 出現場所に関するメタデータで置き換えることにより, 文書画像に特化した圧縮技術を実現している. さらに, 同時に代表擬似コードも保持するため, 圧縮文書画像に対する検索も可能とした. また, 既存の文書画像圧縮技術との圧縮率の比較実験, 検索精度の評価も行った.

次に第 5 章では, 活字印刷文書画像を対象に, 一般のテキスト文書を対象に,M-tree による検索用の索引を利用した高速な全文検索手法について提案する. トランスメディアによる全文検索の従来手法は, 逐次アクセス型の検索であり, テキスト長に対して線形の計算時間を必要とする. このことは大規模な文書集合を対象にしたときに問題となる. 提案する索引技術は一般のテキスト文書に N-gram 索引の手法に着想を得て, トランスメディアにおける擬似コードの列で表現される文書に拡張したものである. また, 活字印刷文書画像を対象に検索評価実験を行い, 提案手法による検索の高速化と精度についても示す. これにより, 提案手法がページ数の増加に対しての計算時間をおおよそ対数オーダーに抑えられることを確認した.

最後に第 6 章では, 続け文字や崩し文字によって形状変動が大きい文字列で構成された, 日本語の手書き草書体文書に代表される古文書に対して, 高速かつ高精度な全文検索を実現する技術について提案している. 草書体文書に適した画像特徴量の提案を行い, さらに, 特徴量値の分布に基づくスカラー量子化を用いたより高精度な擬似コード化の手法を新たに提案した. 提案手法ではまず, 文書画像中の文字領域を含んだ等しい大きさの矩形領域に分割し, 各領域から文字の画像特徴量を抽出する. この特徴量に基づいた擬似コードを各領域に付加することによって, 通常のテキスト文書と同等の文字列検索を画像上で実現する. また, この際に利用する特徴量と擬似コード化手法について, 既存技術との比較のため, 精度とデータ記述量, 文字列照合の計算量の観点から, 実際の手書きの草書体古文書を対象にした評価実験と考察を行った.

本論文では, 印刷, および手書きの文書画像を対象にして, 文字の画像特徴量に基づいた擬似コードを文書の内部表現として用いることで, 文書画像の効率的な圧縮, および, 高速な全文検索を実現する手法を提案した. これらはフォント, 言語に非依存であり, 統一的な手法で適用可能なため, これまで大規模な電子図書館にて活用されてこなかった大量の書籍や古文書の全文検索を用いた利用を可能とする.

学位論文審査の要旨

主　査　教　授　田中　譲
副　査　教　授　原口　誠
副　査　教　授　有村博紀

学位論文題名

A Full-Text Search System for Document Images Based on Character Shape Features

(文字の形状特徴量を利用した文書画像の全文検索技術)

近年, 国内外で大規模な図書館や公文書館の文書の電子化プロジェクトが盛んに進められ, ウェブを通じて利用可能な電子図書館サービスが注目されている. 知的財産権の消滅した文書の全文や, 歴史的な貴重文献や資料が, 画像として広く公開されるようになった. 膨大な文書の中から目的の文書を探すには検索機能が不可欠であるが, 文書の全文検索は, 機械可読な文書か, 文字認識 (OCR) が適用可能な字体で活字印刷された文書画像にのみ適用可能である. 活版印刷時代の文書や, 旧字体の活字印刷文書, 手書き古文書などは, 経年劣化や字体の多様性のために OCR が適用できず, 全文検索の実現が困難であった. 崩し文字や続け文字の場合, 正確に文字を切り出すこと自体が困難である. ウェブを介した文書画像サービスでは, ネットワーク負荷の低減も重要であり, ファイルサイズを一層低減する圧縮技術の開発も望まれている.

本論文は, 活字印刷文書画像および手書き文書画像を対象に, 文字画像の形状特徴に基づいた擬似的文字コードを定義し, これを文書の内部表現として用いることにより,(1) 活字印刷文書画像の高度圧縮,(2) 活字印刷文書画像の高速全文検索のためのインデクシング,(3) 草書体を含む手書き文書画像の全文検索の 3 点に関して著者が新しく研究開発した技術をまとめたもので,(1) に関しては, 圧縮後の全文検索可能性を保証しつつ従来技術に比して圧縮率を著しく改善し, 従来実現が困難と考えられていた (2),(3) に関してはこれらを可能にすることに成功している.

本研究における成果は以下の 3 点である:

1. 活字印刷文書画像に適した画像圧縮手法と, 圧縮ファイルに対する全文検索技術の直接適用手法の提案
2. 活字印刷文書画像に対する M-tree 索引構造を用いた索引付け技術と, この索引を用いた高速な文書画像検索技術の提案
3. 草書体文書を含む手書き古文書画像の全文検索技術と, HMM 学習を用いた適合性フィードバックによる検索精度向上手法の提案

本論文の構成は以下の通りである. 2 章では, 文書画像検索技術, 文書画像圧縮技術に関する既存の研究について解説すると共に, それらの問題点を列挙している. 3 章では, 本研究の基盤となるトランスメディア技術について解説している. 4 章から 6 章では, 上記の各成果について詳細を説明して

いる.

4 章では, 活字印刷文書画像に特化した画像圧縮手法を提案している. 活字印刷文書画像に対する全文検索技術を利用し, 同一文書画像中における同じ文字の出現をすべて見つけ, 異なる文字ごとに 1 つの代表画像を保持すると共に, 個々の文字の出現は, その文字の擬似コード表現と出現場所座標データの対に置き換えることにより, 文書画像用の画像圧縮法を提案している. 文書中の文字の並びに対応してこれらの文字の擬似コードの列が保持されるので, 解凍することなく圧縮ファイルに対して直接全文検索処理を行うことが可能である. 既存圧縮技術との比較評価実験により, 提案手法が圧縮率と検索精度の両面で優れていることが報告されている.

5 章では, 経年劣化のある活字印刷文書画像を対象に, 全文検索の高速化を実現する N-gram インデクスを M-tree を用いて構築する方法を提案している. 提案インデクシング技術は機械可読テキスト文書に用いられる N-gram インデクシング技術に対応しているが, 文書画像は擬似コード列で表現されており, 検索語の擬似コード列と一定の類似度の範囲にある擬似コード列を探せるようにインデクスを構築する必要があり, このために M-tree が用いられている. 経年劣化のある活字印刷文書画像を対象に評価実験を行い, 提案手法が検索の高速化を実現するだけでなく, 高い検索精度を示すことが報告されている.

6 章では, 日本語の手書き草書体古文書に対して, 高速かつ実用精度の全文検索技術が提案されている. 文書画像中の各列の文字領域は等しい大きさの横長のスリット状矩形領域に分割され, スリットごとに画像特徴量を用いて擬似コードが生成され, 文書は擬似コード列で表わされる. この擬似コード列を用いて文字列検索が行われる. 擬似コード列間のマッチングには, 草書体文字列の縦方向の伸縮性を考慮して, 動的タイムワープ法が適用される. 草書体古文書を対象にした評価実験により, 実用上充分な検索制度が得られることが報告されている.

これを要するに, 著者は, 文字の画像特徴量に基づいた擬似コードを文書の内部表現として用いることにより, 経年劣化のある活字印刷文書画像を対象に, 全文検索可能な効率的文書画像圧縮技術と, 高速全文検索を可能にするインデクシング技術, さらには草書体文書画像にも適用可能な全文検索技術に関する新知見を得たものであり, マルチメディア工学, 電子図書館学, 情報検索工学に対して貢献するところ大なるものがある. よって著者は, 北海道大学博士 (情報科学) の学位を授与される資格あるものと認める.