2L-8

Transmediaにおける 圧縮された文書画像に対する文字列検索

猪村 元　　田中譲
北海道大学情報科学研究科知識メディア研究室

1 はじめに

今日，大量の紙媒体の文書による情報を効率的に管理するためにイメージスキャナ等で文書を文書画像として電子化して扱うことが多くなっている．また，検索など柔軟な利用や保存スペースの節減のために，OCR(Optical Character Reader) などの技術が利用されている．しかし現在の OCR の手法においては，文字・言語種に応じて人手によるテンプレート，モデル作成の必要性や，フォントの影響，誤認識などの問題点が存在する．そこで本研究の基盤となる Transmedia システム [1][2] では，イメージスキャナなどによって取り込んだ活字・手書き [3] の文書画像を対象に，文字や言語に依らずに統一的な手法で電子的に管理・利用する手法を提供することを目的としている．本研究では，Transmedia システムによる文書の管理と流通の促進のために，文書が可読であることを保証する新しい文書画像圧縮技術と，この圧縮された文書画像に対しての非展開での文字列検索手法についても述べる．

2 Transmedia 文書

Transmedia システムにおいては文字を文字としての認識は行わずに図形として扱い，各文字の画像より各文字の図形的な特徴をある尺度によって数値として抽出し，この特徴量をもとにして図形的な特徴の反映されたコードを生成する．この際に文字に付与されるコードを Transmedia コード (以下 TM コード) といい，TM コードを付加した文書を Transmedia 文書と呼ぶ．

活字を対象とした本研究では文字領域を等分割した後，特定の 2 領域間の文字ピクセルの密度比を特徴とするメッシュパターン特徴量と，文字の輪郭線を走査してその傾きの変化量が閾値を超える点の数を特徴量とする，アウトライン特徴量の 2 つを用いている．また，特徴量からのコードの生成には，各特徴量を文書中の全ての文字に対して集計してヒストグラムを構築し，これを各階級に属する文字数に基づいて等分したのち，文字の存在する分割領域によってコードを生成していく．これまでは，1 つの特徴量にのみを用いて TM コードを生成していたが，本研究では 2 つの特徴量を組み合わせてコードを生成することにした．これは，メッシュパターン特徴量は文字ピクセルの分布に基づいた特徴量で得られる特徴量数を多く取れる反面，部分的な形状の変化に不安定であり，アウトライン特徴量は文字の形状を反映した特徴量であることから，相補的な効果が期待できるためである．

これらのコードを各文字の画像に対応づけ, 表示には元画像を用いることによって，コードを利用した検索など，計算機による処理を可能としている．処理の概要を以下に以下に示す．(図 1)

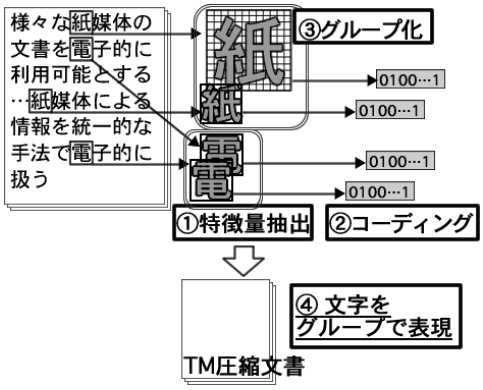

図 1: Transmedia 文書圧縮の流れ

3 文書画像の圧縮

3.1 文書画像の圧縮の手順

Transmedia システムにおける圧縮は，文書画像を文字単位の画像の集合によって表される画像として捉え，同一の文字を表している文字画像をグループ化し，代表画像を用いて表すことにより，通常の画像圧縮の範囲を超えた圧縮を行うことを可能としている．このグループ化は前節で述べた各文字に対応付けられた TM コードを利用して行う．まず第一段階として TM コード間の距離の小さいもの同士をグループ化する．この際の距離尺度にはマンハッタン距離を用いる．コードを用いたマッチングを行うことでグループ化の高速化を図ることが可能となる．そして第二段階として，異なる文字種の文字画像が同じグループに属することがないよう，TM コード

Compressed String Matching on Document Images of Transmedia.
Hajime Imura , Yuzuru Tanaka
Meme Media Laboratory,Hokkaido University
N13W8,kita-ku,Sapporo,060 8628,Japan

によるグループ内の文字画像に対してのみ，実際の画像を用いたテンプレートマッチングを行うことによって精度を高めている．この処理によって生成された各グループにはグループを識別するためのグループＩＤ，グループが表す文字の代表文字画像，そしてグループ内の文字画像に対応づけられた TM コードの平均をとった代表コードが付与される．これによって文書はグループ ID のリストとして表現することが可能となる．さらに，このグループ ID のリストに対して LZW 圧縮を施すことによってさらに圧縮率を上げている．これによって得られる圧縮文書を Transmedia システムにおけるにおける圧縮文書 (以下 TM 圧縮文書) とする．

3.2 圧縮結果

圧縮対象は，朝日新聞の記事の抜粋で，50 ページ，約 65000 字，文字種数約 1600 字，GIF 画像で合計 5.1MB を使用した．比較対照として採用した PDF は，機械可読形式なテキストデータから生成したもので、文書中で使われているフォントとテキストを埋め込んだものを対象とした．結果は Transmedia による圧縮文書が 497kb(1)，PDF が 623kb(2) であった．(図 2) また、Transmedia による圧縮の元画像である GIF 形式の画像サイズも参考に示した (3)。グラフより，文書のサイズが大きくなるにつれて新たに出現する文字が減少するため，Transmedia による圧縮の方はデータサイズの増加が収束していることが見て取れる．

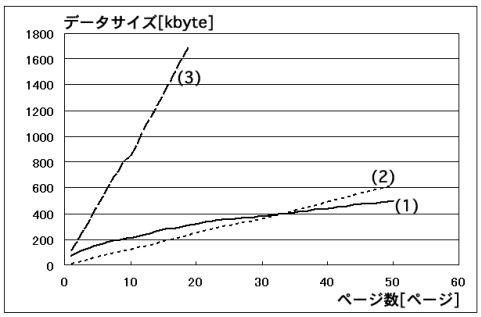

図 2: 圧縮後のデータサイズ比較

4 TM 圧縮文書に対する直接検索

TM 圧縮文書に対する直接検索とは，検索キーワードとして文字列の画像を与え，TM 圧縮文書中から一致する文字列を含む文書画像を検索するものである．そのためにはキーワード画像からこれを構成する各文字のグループの候補を取得し，そのグループ ID を用いて照合を行う必要がある．この照合は LZW 法によって圧縮されたグループ ID 列で表されるテキストに対して，同じくグループ ID の並びで表されるパターンの照合を行うことに相当する．この際の文字列照合アルゴリズムの基盤としては，LZW 辞書を構築すると同時に Aho-Corasick オートマトンにによる照合を行うアルゴリズム [4] を用いている．この TM 圧縮文書に対する検索の詳細を以下に示す．

4.1 キーワード構成文字のグループ ID の取得

TM 圧縮文書中においては 1 つの文字を表すグループが複数存在することがあるため，これを表している可能性のあるグループを全て取得する必要がある．まず，キーワード画像から，2 節で述べた手順により TM コードを生成する．この生成された TM コードをもとに各グループの代表コードに対してコード間の閾値を適切に設定してやることによって，キーワード中の各文字のグループ ID の候補を取得する．

4.2 検索

上記により取得したグループ ID の候補の組み合わせによって表現されるキーワード文字列の出現を一般のテキストに対する文字列照合アルゴリズムによって全て照合することによって直接検索が実現される．ここで，TM 圧縮文書に対して適用するための修正を行う必要がある．文字画像の領域が極端に細長い文字の照合については，メッシュパターン特徴量によって生成される部分のコード間の距離にばらつきが多く見られるため，アウトライン特徴量より生成されるコードのみで照合を行うようにする．

5 まとめ

本稿では，Transmedia システムを基盤とした，文書画像圧縮技術と圧縮された文書に対する非展開の文字列検索の手法について述べた．本システムにおいては，文字を特定の 1 つのコードに変換するのではなく，図形的な特徴をもとにした TM コードを対応づけることで，特定の言語やフォントに依存したテンプレートなどを必要としない．このため，多様な言語・フォントで書かれた文書に対しても検索や圧縮などの処理が適用可能である．

参考文献

[1] 岡田亮，文書画像の圧縮と圧縮データの直接的検索に関する研究， 2000 年度修士論文
[2] 松井啓太，他言語・低品質および手書き文書画像の検索・圧縮に関する研究，2002 年度修士論文
[3] 大橋栄介，手書き文書画像の検索に関する研究，2002 年度卒業論文
[4] T.Kida, M.Takeda, A.Shinohara, M.Miyazaki, and S.Arikawa, Multiple pattern matching in LZW compressed text.