大規模マルチメディア情報の構造化技術

Structuring Technologies for Large-Volume Multimedia Information

堀　修 O. Hori　石谷 康人 Y. Ishitani　清野 和司 K. Seino

大容量のマルチメディア情報がオフィスにも家庭にもあふれつつある。一般に情報洪水と呼ばれるこの問題を解決するために，従来の紙やテープといったメディア上の文書／画像／映像といったアナログ情報をディジタル情報へ変換し，計算機の中に収納する技術が重要となる。その場合，後から情報を検索しやすいように構造化し見出しをつける必要がある。当社では，文書の領域を見出しや本文などのパートに自動的に分類し，文字コードに変換する技術および映像から場面の変わり目を自動的に検出し，シーン単位にまとめる技術を開発した。これにより，人手をかけずに自動的に大規模なマルチメディア情報を構造化し所望の場所が容易に検索できるようになった。

Large volumes of multimedia information are pouring into the home and the office. In order to solve the problem of this so-called "information flood", it is important to digitize analog data such as paper documents, photos, and videotapes, and to store them in computers. This requires the structuring and indexing of data for easy retrieval.

We have developed a document analysis technology for extracting the meaningful parts of documents and changing them into character codes, and a video analysis technology for extracting scene changes and merging them into scenes. These technologies allow the easy retrieval of large-volume multimedia information.

1 まえがき

大規模なマルチメディア情報を扱うための十分な環境が整いつつある。計算機能力の急速な進歩，大容量のハードディスクおよび高速なネットワークの普及により，文字／画像／映像という多種多様なマルチメディア データを操作することが容易になった。

しかるに，文字／画像／映像データの多くの情報は，紙やテープという旧来のメディアの上に載った情報のまま，図書館などのライブラリに保存されている。オフィスへのOA 機器の普及にもかかわらず紙のドキュメントがなくならなかったように，今後も旧来からのメディアが消失することはなく，随時これらのアナログメディアからディジタルメディアへの変換が必要となる。また，単にこれらをディジタルメディアに変換するだけでは不十分で，増え続けるマルチメディア データの情報洪水を避けるためには，それらのデータを構造化し見出しを付けて欲しい部分だけ効率よく検索できる形にする必要がある（図 1）。

ここでは，これらの問題を解決するための基盤技術として，第 2 章で文字／図／写真を含む文書画像の構造解析，第 3 章で映像の構造解析について述べる。

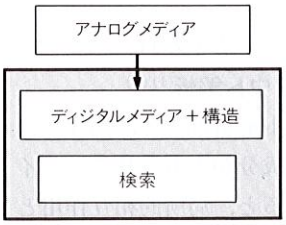

図 1．メディア変換の概念　アナログメディアをディジタルメディアに変換し，検索が容易になるように見出しを付けて構造化する。
Concept of media conversion

2 文書画像の構造化

当社では新聞，雑誌，科学技術文献，ビジネス文書など多様かつ大量の印刷文書を効率よくディジタル化するドキュメントリーダ ExpressReader$_{TM}$ Pro を商品化している。

このシステムでは文書を画像として取り扱い，そこから構造化されたレイアウト情報として①テキスト，写真，図形，表領域，②テキスト領域からタイトル，パラグラフなどのレイアウトオブジェクトと文字行，③文字行から各文字とそのコード情報を抽出し，それらを階層的に記述する。このような構造化により，紙文書から直接，大規模文書データベース（例えば，イントラネットやディジタル図書館）を自動的に構築したり，翻訳・ワープロ・表計算ソフトウェアを利用することが可能となる。

以下ではドキュメントリーダで実現されているレイアウト解析について述べる。

2.1 創発的計算

日本語文書では，①写真／絵／図形などが混在していたり，②縦書きおよび横書き文章（数式，英文を含む）が混在していたり，③領域が近接または入り組んでいることが少

なくない。一般に，多種多様な文書を許容しながら複雑なレイアウト構造を高精度に抽出することはたいへん難しく，従来の製品や研究ではこの問題は解決されていない。ドキュメントリーダでは，このような困難を解決するために創発的計算（Emergent Computation）を導入している。創発（Emergence）とは最近注目されている人工生命研究における重要な概念の一つであり，図2に示すように，「要素間の局所的な相互作用の結果，全体が現れ，その全体が局所的な要素の環境として働き，それにより新たな秩序が形成される現象や考えかた」であるとされている。

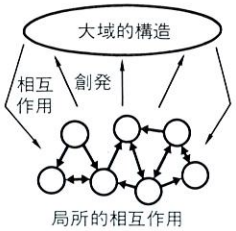

図2．創発の概念　要素間の局所的な相互作用により，そこには見られない大域的な構造が生じ，それがさらに下位の動作に影響を与える。
Concept sketch for emergence

創発に基づく計算モデルを導入することで，定義困難な問題や未知のケースを自己適合的に解決する人工システムの実現が期待される。以下では，一例として創発的計算に基づくレイアウト解析について説明する（図3）。

2.2　創発的計算に基づくレイアウト解析[1),(2)]

ドキュメントリーダのレイアウト解析部には，図3(b)に示す入力画像から抽出された文字成分が入力される。レイアウト解析を図2に対応づけてみると，局所的相互作用として“近接する文字成分の統合”が行われ，その結果として図3の(e)，(f)に示す“レイアウト構造”が大域的構造として創発されると考えることができる。

図3(b)の各文字成分には，その統合可能範囲として一定の小さな値が初期設定されている。最初は統合可能範囲の重なる文字成分間でだけ統合が起こり，小規模な部分領域が生ずる（図3(c)）。ある程度領域が形成されると，各領域で文字行方向，文字行，平均的な文字サイズ，字間，行間などのパラメータが抽出される。このあと，部分領域の統合は，これらのパラメータの影響を受けるようになる。例えば，横書きの部分領域は，横書きで文字サイズがほぼ等しく，水平方向に字間以内あるいは垂直方向に行間以内の距離で近接している他の領域とだけ統合するようになる。

この結果，文字配置の均質さと規則性に基づいた統合が行われ，図3(d)のようにまとまった領域がいくつか生ずる。しかし，この時点では図4(a)のように文字が不規則でスパースに配置されている場合や数式部では統合が生ぜず，正しい領域が形成されていない。

しばらくすると各領域では，その秩序性（まとまりの良さ）が評価されて，“多くの文字行で構成されており，大きな面積を占める高秩序領域”と，そうでない“低秩序領域”に分類されるようになる。この結果，部分領域の統合は周囲の領域の秩序性の影響を受けるようになる。すなわち，高秩序領域が周囲の統合を誘発したり，孤立した低秩序領域の統合が促進されるようになる。

例えば図4(a)のような場合では，二つの低秩序な領域は直下の高秩序な領域と相互作用を行うことで統合範囲が水平方向に広がっていき，やがて正しいタイトル領域を形成する。また，図4(b)のように，参考文献の番号部で文字行方向の推定を誤っている場合には，隣接する高秩序な参考文献の本文部と相互作用することで影響を受け，その文字行方向誤りが解消されて正しい領域が抽出される。

以上のように，隣接する文字成分および部分領域の統合という局所的相互作用は，徐々に成長した領域の幾何的な性質や秩序性という創発された大域的構造からの影響を受

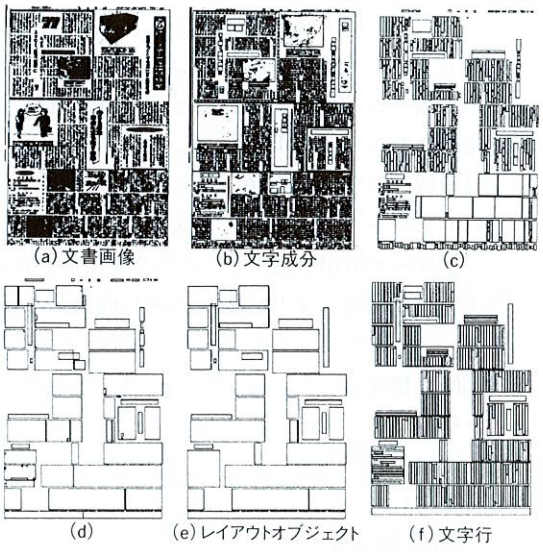

図3．創発的計算に基づくレイアウト解析　近接する文字成分間の局所的な統合により大域的なレイアウト構造が創発される。
Layout analysis based on emergent computation

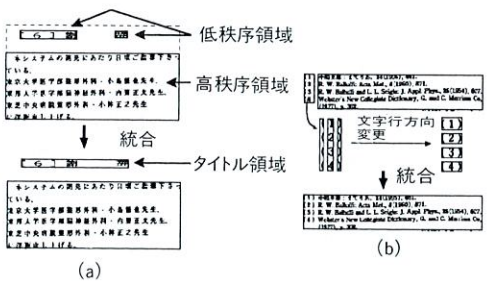

図4．レイアウト解析における上下間の相互作用　局所的統合が周囲に分布する領域の幾何構造や秩序性の影響を受けている。
Interactive computation between global and local layout structure

けていることがわかる。このような上下間の相互作用の結果，入力文書のレイアウト構造の複雑さに適応した解析が可能になる。この結果，これまで解析が困難とされていた，横書き／縦書き文章と写真・図形と数式などが混在し，入り組んでいるような複雑なケースを高精度に解析できるようになった。また，同一の原理により単純な構造のビジネスレターから多段組，多記事の複雑な構造をもつ新聞まで広範囲な文書を一括して取り扱うことが可能となった。

3 映像の構造化

今後，増え続ける多くの映像コンテンツ（テープ）から所望の映像シーンを検索し，効率良く鑑賞する手段の実現が重要な課題となる。ここでは，効率良く映像を見るための映像構造化技術として，カット検出および多数のカットから意味のあるまとまりであるシーンを抽出する方法について述べる。

3.1 MPEG 映像を利用した高速カット検出[3]

カットとは，映像の場面が急に変化する場所で，カットとカットの間はショットと呼ばれる。カットは映像を構造化する上で重要な情報となる。標準圧縮形式である MPEG（Moving Picture Experts Group）は，今後主流になると予想され，MPEG により符号化された映像を完全に復号することなく扱うことにより，高速なカット検出を開発した。MPEG は動き補償付き画像間予測と DCT（Discrete Cosine Transform）のハイブリッド符号化方式である。符号化に際しては，マクロブロックと呼ばれる 16×16 画素サイズの領域に画像を分割し，このマクロブロックごとに画像間予測を行っている。動き補償付きであるため，符号化に際してはマクロブロックごとに参照画像との位置ずれを表す動きベクトルが情報として付与されている。

動きベクトルの符号化に際しては，直前のマクロブロックの動きベクトルとの差分をとり，差分値だけを可変長符号化している。差分値に割り当てられる符号長は出現頻度を考慮して，小さな差分値に短い符号長が割り当てられている。また，予測がうまく当てはまらない場合には画像間予測を行わずに符号化するため，動きベクトルは符号化されない。動きベクトルの符号量とカットとの関係を考察すると，①カットがない場合には画像間の予測符号化が使われ，しかも隣り合うマクロブロックは類似した動きベクトルをもっているため，動きベクトルの符号量は小さくなる，②カットによりまったく異なる画像に変わると，カットを越えた画像間の予測が使われず，動きベクトルが符号化されなくなる，③カット後も同じような色調であるためにカットを越えて画像間予測が行われる場合には，動きベクトルは不揃いになるので，動きベクトルの符号量は多くなる，という関係がある。したがって，動きベクトルの符号量と画像間の類似度とは反比例の関係にある。MPEG 符号化された映像に対し，類似度を求めた例を図 5 に示す。類似度が急激に小さくなっているところがカットである。

カット検出処理を実際に MPEG2 で符号化された映像に対して適用した実験結果を表 1 に示す。ここで用いた映像は 50 分間の映画である。

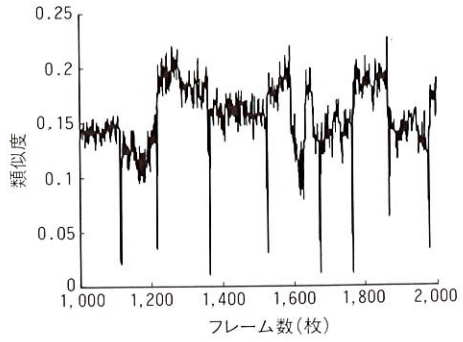

図 5．フレーム間の類似度の例　　カットの部分でフレーム間の類似度が低い。
Similarity of neighboring frames

表 1．カット検出の実験結果
Results of cut detection

総フレーム数	90,000 フレーム
総カット数	521 カット
未抽出数	7/521
過剰抽出数	11/521
処理時間	2,781.4 秒（32.4 フレーム／秒）

3.2 類似ショット統合によるシーンの抽出[4]

カットは約 5 秒に一度の割合で出現することから，2 時間の映画に対しては 1,440 のショットが存在することになる。このショットを複数まとめることにより意味のあるシーンを抽出する。とりわけ映画においては，対話によって構成されているシーンが多い。また，対話シーンだけでなく，二つのショットを交互に繰り返す手法で映像を構成することがよくある。調査によると，ある映画の 50 分間の中に人物が登場しているシーンは 43 分間（87 %）あり，その人物登場シーン中で対話が行われている部分は 39 分間（89 %）であった。類似の映像が繰り返し現われるという特徴に着目し，その区間を自動検出し，ひとまとめにする方法でシーンを検出する。

同じカメラアングルから同じ場面を撮影した二つのショットがあり（同一ショットと呼ぶ），二つのショットの間に別のショットが挿入されている場合，分断された後ろのショットは，前のショットの最終画面から再開する傾向にある。したがって，二つのショットが同一ショットかを判定

するには，時間的に前にあるショットからは末尾のフレームを時間的に後にあるショットからは先頭のフレームを，順に辿っていき（図6），あらかじめ設定した「ショット端からの深さ」までにそれぞれのフレームの組合せで評価し，類似フレームが一つでも存在した場合には二つのショットは同一ショットとする。

二つのフレームの比較には，色相ヒストグラムと輝度分布を用いる。色相ヒストグラムは，画素値が輝度と色相（Y, C_b, C_r）によって表されている際に，$(C_b, C_r)=(\rho\cos\theta, \rho\sin\theta)$という極座標表示を行い，それを$\theta$方向に分割した度数分布である。これは，映像中の移動物体の影響を受けにくく，また映像全体の輝度変化（フェードイン／フェードアウト）の影響も受けにくいという特長をもつ。また，画像を9×9のブロックに分割し，画像全体で輝度分布を正規化して二つのフレームの対応ブロックで輝度変化があったかどうかを調べる。ブロック化することで，画像内の物体移動などの影響を少なくし，輝度分布の正規化で，フェードなどの画像全体の明るさの変化が非類似の判定につながることを防ぐ。二つの比較でフレームが類似した場合，二つのショットは同一ショットと判定しグループ化する。以上の処理でグループ化されたショットは，複数のグループがかみ合わさって場面を構成している。図7に示すようにショットはくし状にかみ合わさって構成され，その部分をシーンとする。

映画“グリーンカード”の50分（532ショット）分に対して実験を行ったところ，同一ショットである353個の繰返しパターン（実際に目視で判定した）に対して，過検出0という条件で340パターン（96.3％）を正しく同一ショットと判定できた。また，これによりシーンを抽出し，ショットごとの532枚のインデックスを300枚（56.4％）に削減できた。この映画中で検出された最長のシーンは4分9秒で，このシーンの中には56ショットが含まれていた。検出されたシーン情報に基づいて映像にインデックスを付け，映像検索が容易になった。

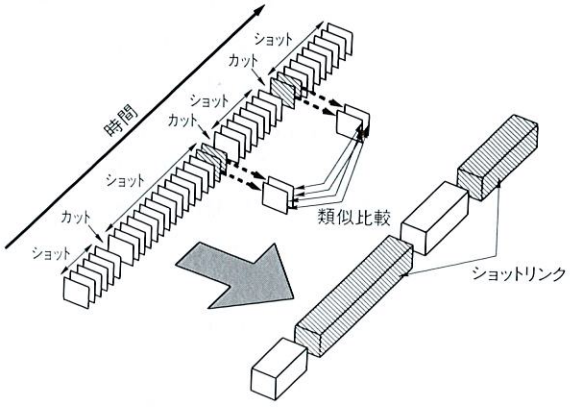

図6．ショットの比較　二つのショットが同一かどうかを検証する。
Comparison of shots

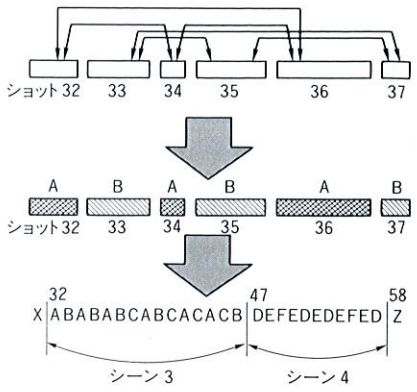

図7．ショットの統合　ショットをグループ化しシーンに統合する。
Grouping of shots

4 あとがき

大規模マルチメディア情報の構造化技術は今後ますます重要になると予想される。ここでは，検索を容易にするための文書画像と映像の構造化技術について述べた。文書処理においては，多種多様な文書に対してつねに安定に構造化できる技術が必要である。映像処理においては，処理速度の向上およびさらに上位の意味にまとめる構造化技術が重要である。

文　献

(1) 石谷康人，他：多階層構造と階層間相互作用に基づく文書構造解析，電子通信情報学会技報，PRMU96-169，pp.69-76（1997）
(2) 石谷康人：創発的計算に基づく文書画像のレイアウト解析，画像の認識・理解シンポジウム MIRU96，I，pp.343-348（1996）
(3) 金子敏充，他：動きベクトル符号量を用いたMPEG動画像からの高速カット検出，電子通信情報学会技報，PRMU96-100，pp.55-62（1996）
(4) 青木　恒，他：映像構造を利用した代表フレーム表示，インターラクション'97，pp.9-16（1997）

堀　修　Osamu Hori
研究開発センター 情報・通信システム研究所研究主務。
文書画像処理および映像理解の研究に従事。
Communication & Information Systems Research Labs.

石谷　康人　Yasuto Ishitani
研究開発センター 情報・通信システム研究所研究主務。
文書画像処理および文字認識の研究に従事。
Communication & Information Systems Research Labs.

清野　和司　Kazushi Seino
青梅工場 コンピュータマルチメディア設計部主査。
OCR認識技術の開発設計に従事。
Ome Works