平成22年5月15日発行（隔月奇数月15日発行）5月号　No.121
ISSN 0914-7535

NHK技研 R&D

コンテンツ活用技術 特集号

2010 No.121 5
日本放送協会 放送技術研究所

巻頭言

コンテンツ活用技術への期待

長谷山　美紀
北海道大学大学院情報科学研究科教授

小さな頃見ていた『ひょっこりひょうたん島』を思い出す。「波をチャプチャプチャプチャプかきわけて」で始まる主題歌は，今も歌うことができる。火山爆発で，突然動き始めた「ひょうたん島」は大海をさまようことになる。遠足で「ひょうたん島」に来ていた子供たちと先生，そして不思議な住人たちが，流れ着く先々で問題に直面しながらも皆で解決していく。もちろん，物語の詳細は時の経過とともに忘れてしまった。それでも，『ひょっこりひょうたん島』を見て楽しく過ごした時間は今でも鮮明に覚えている。

『ひょっこりひょうたん島』から40年あまりたって，通信と放送の融合の議論が起こり，現在は，放送と通信の連携サービスのための技術開発が行われ，幾つかのサービスが開始されている。視聴者の多様なニーズに応えるためにも，また，新規ビジネスフィールドを創出するためにも，コンテンツ活用技術の開発は社会的重要性を増している。

地上波デジタルテレビ放送が全国で開始され，テレビ視聴機器は「放送」と「通信」双方の機能を有する新しいハードウエアに移行している。一方，平成21年版情報通信白書によると，平成20年末のインターネット利用者数は対前年比3.2%の増加であり，その率は小さいが，現状も増加傾向にある。更に，自宅のパソコンを使ってインターネットを利用する人の86.9%がブロードバンドを利用していることが報告され，インターネットの利用目的についての調査では，「平成19年末から最も利用が伸びたものはデジタルコンテンツ（音楽・音声，映像，ゲームソフト等）の入手・聴取」との報告もなされている。視聴機器の変化と高速な通信基盤の普及は利用者のコンテンツ視聴を更に促すものと考える。

ここで，再度，認識すべきことは放送コンテンツとWebコンテンツの差異である。常に述べられることではあるが，放送コンテンツは高品質であり，信頼性も高く，皆が安心してそれを楽しむことができる。Webコンテンツには，高度な技術を持ち合わせたクリエーターが作成している芸術性の高いものもあれば，一般の人が作成しているものもあり，その品質は多様である。また，配信ビジネスの形態によっては，信頼性の高い映像もあるが，一般には，信頼性は保証されていない。放送コンテンツの品質と信頼性は特筆すべき性質である。利用者のコンテンツ視聴が促される環境で，品質と信頼性が約束された放送コンテンツの活用が進めば，利用者が豊富な情報に触れ，知識を獲得するための社会基盤が形成されると考える。これは，放送コンテンツが社会に果たす重要な役割である。

更に，具体的に開発される技術に目を向けてみよう。放送コンテンツを活用するためには，利用者が希望の映像を手に入れるための検索エンジンが必要である。検索エンジンはWeb上の大量かつ多種多様な情報から必要な情報を取り出すために使用されてきたが，放送コンテンツの活用においても同様のニーズが発生する。映像の検索エンジンを実現するためには，映像内容を説明する情報であるメタデータを効率よく付与する技術が用いられる。検索エンジンは付与されたメタデータ同士を比較照合して，利用者が望む映像を提示する。利用者が満足する映像を提示するためには，メタデータを高精度に抽出する必要があり，盛んに研究が行われている。映像コンテンツに対す

1988年　北海道大学大学院工学研究科修士課程修了
1989年　北海道大学　応用電気研究所　助手
1994年　北海道大学工学部　助教授
2006年　北海道大学大学院情報科学研究科　教授
現在に至る。

1995年から1996年まで，米国ワシントン大学客員准教授。現在，画像・映像および音響信号などマルチメディア信号処理の研究に従事。博士（工学）。総務省総情報通信政策局情報通信審議会専門委員。経済産業省「ITとサービスの融合による新市場創出促進事業プロジェクト」戦略委員。

るメタデータの抽出では，単独の手法では十分な精度を得ることが難しく，映像認識，音声認識，言語解析などさまざまな分野の複数の手法を用いて抽出精度を高める方法が有効とされている。そのため，手法を自由に組み合わせて利用できる研究の共通基盤も開発され，仕様や抽出モジュールのサンプルプログラムも公開されている。連携による更なる技術の高度化が期待される。

このようにして高精度なメタデータが得られ，それらを照合して検索エンジンが利用者に希望の映像を提示しようとすると，まず，利用者が検索要求をキーワードなどで表現する必要がある。そのため，キーワードを具体的に想像することが困難な場合においては，ユーザーが望む映像を視聴できないという問題が生じる。キーワードを明示することは，検索エンジンの利用に慣れている計算機の利用者でさえも難しい。どのような利用者も映像を見たいと思ったときに，自分自身の検索要求をキーワードで適切に表現できるとは限らず，一度の検索で目的の映像にたどり着くことができなければ，繰り返し検索を行わなければならない。このような検索形態が，放送コンテンツの視聴に適しているかと言うと，返答に苦しむところである。既に，たくさんの利用者が存在する放送コンテンツでは，その視聴において，過去に利用者にこのような作業負担を強いたことは無かったと思う。今までのテレビのように簡単に視聴ができる環境を提供しながら，大量のコンテンツから利用者が望む映像を手に入れるためのシステムが必要である。もはやこれは，検索と言う言葉の定義を超えているように思う。利用者の好みに合わせて，システム自身が検索だけでもなく，推薦とも言えない，独自の融合形態を形成することで実現される，気づきの支援者のようにさえ見える。このように考えると，大量に保持された放送コンテンツを活用すると言うことは，放送コンテンツの豊富な情報を通して，すべての利用者が知識と感動を享受するための「メカニズム」を実現することかもしれない。

本特集号は，その実現について示唆を与えてくれる。画像から自動的に意味的内容を抽出する情報解析技術の動向が示され，各種技術の連携利用による高度化を実現するメタデータ制作フレームワークについて解説が行われ，更には，付与されたメタデータを活用したコンテンツの自動検索・推薦システム，言い換えると，映像コンテンツの活用を図る新しい視聴スタイルの構成例が紹介されている。放送コンテンツの活用にふさわしい「メカニズム」が見えてくるように思う。

放送コンテンツは私たちに知識と感動を与えてくれる。その社会的役割は極めて重要である。『ひょっこりひょうたん島』の主題歌では，「丸い地球の水平線に何かがきっと待っている」と歌われている。映像コンテンツに内在する知識と感動を利用者が享受する「メカニズム」が実現されるとき，その先に何が待っているのだろうか。生み出される未来に期待したい。

解説

コンテンツ活用技術の概要

藤井真人　　柴田正啓

放送と通信が連携した仕組みを利用して，放送局が制作し保有している大量の映像コンテンツを視聴者や利用者が使いやすい方法で提供することが求められている。このためには，映像コンテンツが意味内容に基づいて検索できることが基本となる。本稿では，映像コンテンツを活用して新しいサービスを展開していくための内容解析や検索・推薦の技術について，その考え方と研究開発の動向について概観する。

1．まえがき

「コンテンツの創造，保護及び活用の促進に関する法律」（通称，コンテンツ保護法）では，コンテンツとは「映画，音楽，演劇，文芸，写真，漫画，アニメーション，コンピュータゲームその他の文字，図形，色彩，音声，動作若しくは映像若しくはこれらを組み合わせたものまたはこれらに係る情報を電子計算機を介して提供するためのプログラム（電子計算機に対する指令であって，一の結果を得ることができるように組み合わせたものをいう。）であって，人間の創造的活動により生み出されるもののうち，教養または娯楽の範囲に属するものをいう」と規定している[1)]。この定義によれば，コンテンツという概念には人間の創造的活動によって生み出された教養，娯楽の範ちゅうに属する成果物一般を包含し，コンピューターを介して成果物を提供するためのプログラムも含めるということになる。人間の創造的活動によって生み出された成果物は従来から「作品」などと呼ばれてきたが，上記のように「コンテンツ」という言葉を使う場合には，コンピューターによって提供されるという側面が意識されていることが多い。ここで，コンピューターで処理できるということの前提は情報がデジタル化されているということであり，アナログ時代の「作品」はデジタル化によって「コンテンツ」と呼ばれるようになるとも考えられる。

このことを放送に当てはめると，アナログ時代から提供されている放送番組は放送のデジタル化によってコンテンツの仲間入りをしたということができる。デジタル放送では，映像，音響データを組み合わせることによって構成された番組のほかに，データ放送によって種々の情報が提供されている。また，電波媒体だけではなくWorld Wide Web（以下，Webと呼ぶ）やその他のネットワークを介してニュースや番組本編のほか，さまざまな番組関連情報が提供されている。これらも併せて放送によって提供されるコンテンツという意味で，放送コンテンツと呼ぶことができる。

本稿では，このような意味での放送コンテンツを，より多くの視聴者に，より使いやすい方法で提供するための技術の研究開発について紹介する。この技術では，映像，音

響といった非言語情報が主体となる放送コンテンツに簡単にアクセスし操作するためのメタデータ（データを記述するデータ）が重要である。メタデータをできるだけ人手を掛けない方法で放送コンテンツ自体から抽出する技術，メタデータを使って映像の視聴形態に適した方法で放送コンテンツを検索または推薦する技術などが放送コンテンツをアナログ放送時代にはない方法で活用していくためには必要である。

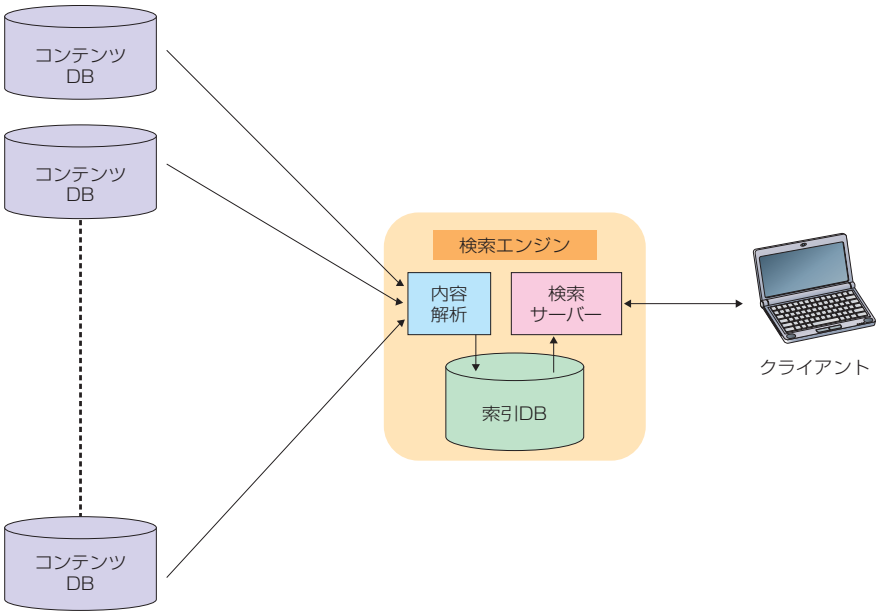
2．コンテンツ活用技術の考え方

2.1　メタデータの役割

コンテンツの活用にとって，最も基本的な技術は，利用者が所望のコンテンツに簡単にアクセスし閲覧することを可能にする検索技術である。今やインターネットを介してコンテンツを配布，閲覧する仕組みであるWebは多くの人々にとって日常的な存在になっている。Webの普及にとって，Web上で稼動する検索エンジンの役割は非常に大きなものであった。

検索エンジンの概念的な構成を**1図**に示す。検索エンジンはコンテンツを集めてその内容を解析し，各コンテンツにアクセスするための情報とその内容を記述し検索の手がかりとなる情報を対応付けた索引（インデックス）を作成し，データベース（DB）として保持する。利用者がクライアント端末から入力した検索要求を受け付けた検索エンジン内の検索サーバーは索引DBを参照して，検索要求に適合したコンテンツへのアクセス情報を回答するというのがおおよその検索の仕組みである。

Webで公開されているコンテンツ（以下，Webコンテンツと呼ぶ）の場合には，1図のコンテンツDBはコンテンツを公開するWebサーバー，コンテンツにアクセスするための情報はURL（Uniform Resource Locator）に相当する。また，Webコンテンツは多くの場合HTML（Hyper Text Markup Language）の規格に従ってタグが付与された自然言語の文章が主体である。検索エンジンはWebコンテンツを収集し，HTMLのタグを参照しながら文章を基本単位である単語に分解し，単語の重要度などを考慮して索引語を



1図　検索エンジンの概念的な構成

選択する。各索引語とその単語を含むWebページのURLの対照表としてインデックスを構成し，索引DBに格納する。利用者の検索要求はキーワードという形式で指定され，検索エンジンはインデックス中の索引語とキーワードを照合することによって，適合するWebページのURLとタイトルなどの情報を検索結果として出力する。この過程では，文章の解析や複数の検索結果間のランク付けなどの処理を行うが，基本的な技術は自然言語処理，情報検索技術として研究開発されてきたものが基盤となっている。これらの技術やネットワーク技術などの進展によって，検索エンジン側にWebコンテンツをWeb全体から効率的に収集し，文章の内容を解析して大規模な索引DBを自動生成し，管理することが可能になった。このことがWebの検索エンジン利用の普及につながっている。

放送コンテンツの場合でもWebで提供される番組関連情報などは文章が主体であり，上記と同じような仕組みの検索エンジンを使って利用者に提供することができる。これに対し，映像，音響といった非言語情報が主体となるコンテンツ（以下，映像コンテンツと呼ぶ）は，デジタル放送として編成された時系列的な流れの中で電波やCATVなどを使って提供されることが一般的である。しかし，近年，専用ネットワークやWebを介したオンデマンド型の提供が可能になってきた。更に，映像アーカイブスが多数構築・公開され，オンデマンド型で提供される映像コンテンツの量が拡大すると，階層分類などを使った単純なコンテンツ選択方法では不十分で，Webコンテンツと同様，あるいは，映像としての特質に合わせたより高度な，内容に基づいた検索の機能が求められるようになる。

検索エンジンで，映像コンテンツを内容に基づいて検索するためには，Webコンテンツの場合と同様に，利用者の検索要求と映像コンテンツの内容を照合するためのインデックスを作成する必要がある。ただし，映像や音響は連続的な時系列データであり，文章の単語に相当する意味的な基本単位が明確ではない。動画像データは静止画データの系列と考えることができるので，個々の静止画は物理的な基本単位とはいえる。しかし，これは必ずしも単語のような意味を担う単位ではない。Webコンテンツの検索では，意味を担う単位としての単語を用いてインデックスを構成し，利用者も検索要求を単語（キーワード）で表現することで，インデックスと検索要求の直接的な照合を行う。仮に，単語の代わりに静止画を用いてインデックスを構成したとしても，利用者が検索要求を静止画で表現することは簡単ではなく，この方法では誰でも簡単に使えるように，内容に基づく映像コンテンツ検索機能を実現することはできない。映像コンテンツの内容に基づく検索を実現するためには，利用者の検索要求と照合可能なレベルの内容記述データが必要である。この内容記述データは人手によって映像コンテンツに付与する，あるいは，内容解析を行って映像コンテンツから自動的に抽出することによって，映像コンテンツの上位の層のデータ，すなわち，メタデータとして作成されるべきものである。

2.2 コンテンツ活用技術の研究

上述のように，映像コンテンツの活用では，まず，メタデータを作成することが重要である。ここでいうメタデータの基本的な役割は，利用者の検索要求と照合可能な映像コンテンツの内容についての索引ということになる。Webコンテンツと同様に，キーワードで表現された検索要求に対応するためには，映像の内容を自然言語のような記号で記述したメタデータが必要である。また，映像コンテンツはそれぞれが意味を持ったシーンの連なりによって内容が展開していくので，内容を記述するメタデータは個々のシーンを切り分ける境界の情報も持つ必要がある。このほか，自然言語では表現しにくいシーン同士の類似性に基づく検索を実現するためには，シーンの映像としての特徴を

うまくとらえる情報が必要になる。これらの情報を，多くの労力を掛けずに効率的に作成する技術を開発することが，コンテンツ活用のための内容解析技術の研究の目的である。

内容解析技術の研究では，映像コンテンツをコンピューターによって解析して，メタデータに必要な内容記述情報を取り出す技術が開発されている。映像，音響データから内容記述情報を取り出すためには，パターン認識技術が必要である。コンピューターの処理能力や記憶容量の向上によって，大量の学習データを用いた機械学習を始め，さまざまな統計的手法がパターン認識のために使えるようになった。これらを駆使して，映像コンテンツを構成する映像，音声，言語データを統合的に解析する技術の研究が進んでいる。

検索エンジンでの内容解析では，収集した大量のコンテンツデータを効率よく解析することが重要である。Webコンテンツの場合にはコンテンツの主体である文章から，基本的には他の情報を参照することなく直接的にインデックスを構成することができる。ここで，インデックスを構成する索引語はコンテンツから切り出した単語であって，コンテンツと同じ層の情報であるという意味で，メタデータとは異なるものと考えられる。Webコンテンツのメタデータは基本的にはコンテンツの作成側で付与し，特定のタグを目印にしてコンテンツに付随させることができる。作者名や権利など，内容解析では取得できない情報を含み，より詳細な条件を指定した検索を行うことが可能である。

これに対し，映像コンテンツの場合にはコンテンツデータの情報量が大きく，検索エンジン側にコンテンツデータを広い範囲から収集して，すべての内容解析を行うことは難しい。従って，大規模で網羅的な検索サービスを実現するためには，コンテンツ提供側でも内容記述情報を含むメタデータを作成し，これを検索エンジンが収集できる仕組みを開発する必要がある。このためには，相互運用性のある標準化されたメタデータを作成・共有・活用できることが重要であり，コンテンツ活用技術の研究では，メタデータの標準化や制作のための基盤およびこの基盤を活用し映像コンテンツに適した検索・推薦や提示を行うための手法が研究開発されている。

以下，これらの動向を概観する。

3．コンテンツ活用技術の動向

3.1　内容解析技術

ここまで述べてきたように，大量の放送コンテンツの中から期待する番組やシーンを検索できるようにするためには，映像の内容を表す情報，いわゆる，メタデータを基に検索する方法が有望である。ここでは，コンピューターを利用した解析により映像から自動的にメタデータを取り出す技術について概観する。映像そのものを解析する手法のほかに，字幕や台本などのテキスト処理を併用する手法も取り上げる。

（1）ショット境界検出

意味内容に基づく映像検索を可能とするためには，意味を持ったシーンごとに映像を分割することが望ましいが，この分割処理は難しい課題である。映像コンテンツ，特に，放送番組や映画などは，通常，多くのショットから構成されている。ここでいうショットとは１台のカメラで連続して撮影された映像区間を指す。映像コンテンツを分割するステップとしては，まず，ショットを基本単位とすることが比較的容易である。コンテンツ解析技術の分野では，研究の早い段階からショット境界の検出手法が検討されてきた。

2001年から始まった映像検索技術評価のワークショップであるTRECVID（TREC Video Retrieval Evaluation）[2]において，ショット境界タスクが2008年まで行われてきた。TRECVIDは米国国立標準技術研究所（NIST）等が後援するワークショップで，毎年４つほどのタスクが決められ，参加者は挑戦するタスクを選び，共通の映像に対して，決められた期間内に自分たちの処理結果を報告する。当所で提案したショット境界の検出手法は，2008年に速度と精度の両面において優れた性能を示した*1。

（２）一般物体認識[3]~[8]

物体名は基本的なメタデータであり，これを自動的に付与するために物体認識技術が使われる。最近，物体上の幾つかの点（Keypoint）の周りの情報だけを用いて物体を認識しようとする研究が注目されている。その代表的なものがSIFT（Scale Invariant Feature Transform）*2と呼ばれる特徴量を使う方法である[9]。SIFTは画像のずれ，拡大・縮小，回転にロバストな特徴量である。また，SIFTと比較して計算コストの低いSURF（Speed−Up Robust Features）*3という特徴量も提案されている[10]。認識の段階では，テキスト処理で利用されている手法が適用される。その手法では，文章を分類するときに文章を構成している単語の組み合わせのみに注目し，分類した文章とその文章中の単語の組の関係を機械学習により習得するものである。この基本的な枠組みを画像認識に当てはめ，例えば，文章を物体に，単語を物体のSIFT特徴量に対応させることで，テキスト処理の手法を適用する。テキスト処理での語順を無視した処理と同様に，Keypointの位置関係を無視した処理となる。そのため，観察条件や物体の動きによりKeypoint間の位置関係が変わっても安定して対象物の識別を行うことができる。この手法は汎用性が高く，昨今のコンピューター能力の向上や機械学習の発展とともに適用範囲を広げている。

（３）人物に関する情報の抽出

○顔画像処理[11][12]

放送番組や映画の映像中に頻繁に現れるものは，なんと言っても人である。そのため，人にかかわる情報をメタデータとして付与する研究は数多く試みられている。顔画像認識を利用した手法がその典型であるが，背景には，デジタルカメラの顔検出機能が実用化されるなど，ここ数年間の顔画像検出技術の急速な進歩がある。一般に画像処理の計算コストは高く，検出精度が高くても実用的な処理時間で動作する技術の開発は難しかった。しかし，Violaら[13]が実用に耐えうる精度と処理時間で動作する手法を提案したことで，顔画像検出技術の利用が急速に進んだ。当所でも，この手法をベースにして，より安定して映像から顔画像が検出できる技術を提案している[14]。

一方，顔画像検出だけでは，誰が映っているのかに関する情報を得ることはできない。顔画像から人名を知るための技術は顔画像認識と呼ばれ，多くの手法が提案されている。顔画像認識技術では，新規人物の顔画像を登録する必要があるが，この処理は一般に煩雑である。そこで，当所では，比較的容易に新規人物の顔画像を登録する手法を検討している[15]。Kumarら[16]は，顔に関するさまざまな情報を検出する判定器（性別，人種，年齢，髪の毛の色，表情など65種類）の出力結果を基に，人物を判定する手法を提案している。この手法では，特定人物の画像のさまざまな変化に柔軟に対応できると報告している。

特定の人物ではなく，多くの人物の映像に対して広く人名を付与するために，手動で新規登場人物の顔情報を登録することには限界がある。そこで，マルチモーダル処理*4により，別の目的で人手によって付けられた文字情報を活用して，登録人物に対する学

＊1
本特集号の報告「逐次的な判定手続きに基づくショット境界の高速検出手法」参照。

＊2
画像の局所的な特徴量。標準偏差の異なる複数の正規分布に基づいてフィルター処理を行い，撮影変化に対して安定な点（Keypoint）を見つけ，その点の周りの輝度変化の情報をブロック位置や輝度変化の方向ごとにヒストグラム化した量。

＊3
SIFTの正規分布の計算法を簡略化・高速化して計算コストを軽減した特徴量。

＊4
種類の異なる複数の情報源（画像，テキスト，音など）を対象とする処理。

習データを集める手法が提案されている。Satoh[17]らは，テレビのニュース映像を対象に，テロップ（オープンキャプション）から得られた人名と顔画像の共起確率を基に人名をメタデータとして付与する手法を提案している。Ramananら[18]は，11年分のテレビ映像資料から約60万人分の人物のデータセットを半自動で作成した。Ramananらの場合には，同一人物に対する年齢の違いや，顔の肉付きの変化も含めた記録が可能であると報告されている。コンテンツが映画の場合には台本も利用できる。台本と字幕のテキスト情報を用い，それらを顔画像検出により得られた顔画像情報へリンクさせ，出演者の出現シーンに役名を自動付与する手法が提案されている[19][20]。Everinghamら[19]の手法では，①台本の出演者ごとのせりふと，字幕の時刻ごとのせりふを比較し，台本の出演者名と字幕の時刻を対応付け，②それらの情報と出演者が話しているときの顔画像検出結果を使って人物を特定し，③顔の映っているシーンに広く人名を付与するために顔と服装のマッチングを併用するといった工夫を行い出演者名をメタデータとして自動的に付与する。Phamら[21]はWebのオンラインのニュース映像を対象に，顔画像検出とキャプションを用いて，複数の人が映っているシーンにも，それぞれの人に人名を付与する手法を提案している。

○人体の検出

顔画像検出は顔が正面を向いているときには高い性能を示すが，正面から向きがずれると検出率は一般的に下がる。また，サングラスや帽子を着けている場合などでは，顔を手がかりに人物を検出することが難しくなる。そこで，身体を検出する手法の検討も進められている。その中で，特に注目されている手法がHOG（Histograms of Oriented Gradients）*5特徴を用いる方法である[22]。HOG特徴量は明るさの変化量のヒストグラムを局所的に作成したもので，おおまかな形状を識別することに適すると言われており，形状の変化が比較的大きい人物の検出に向いている。また，顔画像検出と顔画像検出が成功したときの服の色から，顔画像検出がうまく働かない場合にも，服を識別することで人物を認識する手法も提案されている[16][19]。ただし，この手法の場合には，出演者の服の色が変わらないことを前提としているので，適用するときには注意を要する。

（4）イベント検出

人物の検出や認識の結果は「人がいるシーン」や「○○さん」という検索要求に答えるためのメタデータにはなるが，人がどのような状態で，何をしているかという検索要求には答えられない。また，スポーツ映像に対しては，例えば，野球のホームランやサッカーのシュートなど，各競技における重要な出来事を検索したいというニーズもある。このような動作や出来事のシーンにメタデータを付与するための技術は一般的にイベント検出と呼ばれ，動詞的な内容で検出するためのメタデータを提供する。

○人物の姿勢，動作・行動の抽出

身体の様子の基本的な情報として，姿勢をクエリー*6として検索するための技術が提案されている。Ferrari[23]らは，上半身を頭，胴体，上腕，前腕とに分けて検出することで，体の姿勢を識別する手法を提案した。その手法を用いて，腰に手を当てる，腕組みをするなどの姿勢の検出が試みられている。

動作の検索についても多くの研究報告がある。動作についても人名と同じように，映画を対象に，台本と字幕を使い，画像特徴と関連付けることで動作名を付与する手法が提案されている[24][25]。また，Web上の画像からキャプションを頼りに動作の状態が映っている画像を自動収集し，YouTubeの映像から特定動作を検出する手法も提案されている[26]。特定の動作の認識については，一般物体認識と同様に，言語処理で培われた技術，

＊5
SIFTと同じようにブロックごとの輝度変化のヒストグラムを利用するが，特定の点（Keypoint）ではなく，任意の場所で一定領域の範囲を記述する特徴量。

＊6
検索のための手がかりとする情報。通常は，テキストをクエリーとすることが多い。

例えば，確率的潜在変数モデル*7であるProbabilistic Latent Semantic Analysis（PLSA）を用いる手法や[27]，テンソルの分解*8により認識する手法[28]などが提案されている。動作認識は，認識対象が時間変化の成分を含むうえに，観察される姿勢が向きの違いも含めてさまざまに変化するので，大きく異なる２次元画像の時系列データを比較して，動作を判定しなければならない。そのため，認識の性能向上を目的として，文脈を利用する手法も検討されている[29)~31)]。ここで言う文脈とは，認識対象が存在している空間内にある他の物体や背景情報のことを指す。文脈を使う利点は，例えば，指の動きだけを観察して，ピアノを弾いているシーンか，コンピューターの入力装置のキーボードを打つシーンかを区別することは難しいが，背景にあるピアノの鍵盤やコンピューターが認識されると，各文脈（背景）情報の制約により，ピアノを弾いているのか，キーボードを打っているのかを比較的容易に区別できるであろうという点にある。

TRECVIDでは，2008年からイベント検出として，空港に設置したカメラ映像を対象に，走っている人や携帯電話をかける人などの，特定の動作をしている人を検出するタスクを設定している。当所でも，昨年からこの課題に挑戦しており，2009年にはHOG特徴とサポートベクターマシン（SVM）識別法*9を用いた人物検出と，人物を追跡して得た移動軌跡を分析する手法を用いて「走る」イベントにおいて，１位の成績をあげることができた[32)]。

○スポーツ中継映像からのイベント検出

スポーツ中継映像に対しては，見逃した試合を短時間で視聴したいという要求やスポーツ中継映像の中からイベントを検索したいという要求などがある。これらの要求に答えるために，イベントごとにメタデータを付与することは重要であり，スポーツのイベント検出やダイジェスト生成の研究が進められている[33)34)]。Guptaらは[36)]，字幕を使い野球のイベントの映像推移を自動で学習する手法を提案しており，Storylineと呼ぶ行動とその因果関係を木構造で表して，各イベントのモデルを学習する方法を提示している。当所では，野球中継のカメラの切り替えがある程度は定型化していることに着目し，ショットの並びを手がかりに，ホームランやヒットなどのイベントを検出する方法を提案している*10。また，野球の球種をメタデータとして付与するための技術も提案している[37)]。サッカーは野球と異なり展開の激しいスポーツなので，ショットの切り替わりを定型化しにくい。そこで，選手配置とその動きに基づいて映像の内容を分類する手法を提案している[38)]。

3.2　映像要約技術

たくさんの検索結果の中から，目的のコンテンツかどうかを短時間に見極めるためには，要約映像が役立つ。要約映像には，２種類があると考えている。１つはハイライトシーンを時系列に並べた要約映像であり，他の１つはこれから映像コンテンツを視聴してもらうことを目的とした，コンテンツを紹介する要約映像である。前者の研究は多く行われているが，後者は難しい課題であり研究例は少ない。当所では後者の課題に取り組んでおり，字幕と番組の内容を紹介する文章を使ってテキスト処理を行い，番組紹介映像を自動生成する手法[39)]を提案している*11。

3.3　物理特徴による映像検索技術

画像の中に映っているものや音楽の題名を知らなくても，画像の雰囲気や音楽の印象を手がかりに画像や音楽を検索したい場合がある。このような検索意図に答える目的で，画像の色や模様，大まかな構図を手がかりに検索する手法[40)~42)]や，音楽のメロディーやリズムの類似性に基づいて検索する手法の研究が行われている[43)]。また，類似画像検索

＊7
テキストや画像の生成過程を考えるときに，観察されない変数（潜在変数）を経由してこれらが確率的に生成されるとするモデル。例えば，文章のトピックや画像のカテゴリーなどの潜在変数によってテキストや画像が確率的に生成されるとするモデル。

＊8
ベクトルの概念を数学的に拡張したテンソル（多次元配列）の各配列に，異なる特徴を割り当て，行列の特異値分解と同様な分解処理を行って主となる特徴量を抽出する手法。

＊9
多くの正解データを使って識別方法を自動的に獲得する機械学習手法の１つで，基本的に２値の識別関数（顔か顔でないかなどの判定手段）を得る手法。未知のデータに対して高い識別性能を示す。

＊10
本特集号の報告「投球の次ショットに重きを置いたシーンのシンボル列化による野球放送映像プレイ種分類」参照。

＊11
本特集号の報告「電子番組表における紹介テキストを利用した番組紹介映像の自動生成」参照。

のクエリーとしてスケッチを用いる手法も提案されており，当所では，検索対象となる映像群によく現れる画像小片をパレット上に並べ，それらの組み合わせで類似画像検索のクエリーを作成する仕組みを提案している[40]。また，画像合成に類似画像検索を利用した方法も提案されており，画面上にラフな線画と線画で描いた物体の名前を入力すると，各物体を含む画像が自動検索され，その中から画像合成しやすい絵柄の画像の組み合わせが選ばれ，自然な合成画像が実現できる[42]。

このほか，同一場面を異なる視点や時間で撮影した映像を検索する目的で映像解析をする技術の研究も進められており，例えば，撮影対象の動きに着目し，特徴点の動きの軌跡の時間的変化に基づき，同一場面の映像を検出する方法が提案されている[44]。

3.4　メタデータ制作システム

実際のサービスに活用するメタデータでは，高い信頼度が要求されるので，コンテンツ解析技術による自動抽出機能を活用する場合においても，人手による編集を要することが多い。また，権利処理などを含む書誌的事項*12はコンテンツ解析によって抽出できるものではなく，これらをメタデータに含める場合には，編集作業が必須になる。

この編集作業を効率的に進めるために，複数の自動特徴解析機能やエディター，DBを連携動作させる構成のメタデータ制作システムが欧州の共同研究プロジェクトなどで開発されている。MUSCLEプロジェクトの４M[45]，aceMediaプロジェクトのM−OntoMat−Anotizer[46]，PrestoSpaceプロジェクトのMAD[47]，Fraunhofer研究所のiFinder[48]などがある。これらのうちの幾つかはソフトウエアが公開されている。また，プロジェクトの中で異なる機関の開発するサブシステム間の連携を可能にする仕組みも作られている。当所では，このような考え方を更にオープンにしたメタデータを制作するためのフレームワーク（Metadata Production Framework：MPF）を提案し*13，仕様と複数の自動解析機能を統合しメタデータの編集が可能なメタデータエディターを始めとするサンプルのソフトウエアを公開している[49]。

3.5　映像コンテンツの推薦・提示技術

当所では，検索を活用して，テレビ視聴中にシステムが自動的にコンテンツを推薦する仕組みも検討している*14。これは，視聴中の番組やシーンに付与されたメタデータを使い，検索対象のコンテンツのメタデータと比較照合することにより，関連するコンテンツを検索し推薦するものである。メタデータ，検索システムと表示システムのインターフェースなどの仕様を決めることで，その仕様に基づいてさえいればサービス提供側システムと表示端末を別々に設計できるようにしている。また，コンテンツを活用した新しいサービスを開拓するために，映像を中心としたマルチメディア百科事典の自動構築の研究[50]や，画像解析，自然言語解析を使い自動的にクイズコンテンツを生成する方法の研究も行っている*15。

4．あとがき

検索を目的とした映像解析技術の研究は幅広く，本稿で取り上げた内容はその一部であることをお断りしておく。

ところで，TRECVIDにおいては，物体名やイベント名を検出するタスクとして高次特徴抽出課題が設定されているが，参加者中最も良い成績のものでも，実用レベルにはかなりの隔たりがある。セマンティックギャップと呼ばれている物理量から意味や概念レベルの内容を取り出す技術の難しさを示す結果となっている。映像解析により自動的にメタデータを付与する技術は，本稿で取り上げたようなアイデアに富む魅力的な手法が

＊12
制作者，放送日，権利処理などコンテンツの制作や配信などに関する情報。

＊13
本特集号の解説「メタデータ制作フレームワーク」参照。

＊14
本特集号の解説「コンテンツを自動的に推薦するテレビ」参照。

＊15
本特集号の報告「蓄積されたニュース番組からの画像付きクイズ生成手法」参照。

数多く提案されている一方で，セマンティックギャップを埋めるためには，まだ，着実な技術の進歩を必要としている。検索要求は多様であり，その要求に答えるための抽出技術も多種多様にならざるを得ず，そのためにも，多くの研究者や技術者が協力してメタデータ自動付与技術を開発し提供できる環境の整備は重要であると考えている。

参考文献

1) コンテンツの創造，保護及び活用の促進に関する法律，
http://www.cas.go.jp/jp/hourei/houritu/kontentu.html

2) TRCTVIDのホームページ，
http://wwwnlpir.nist.gov/projects/trecvid/

3) 柳井：“一般物体認識の現状と今後，”情報処理学会論文誌コンピュータビジョンとイメージメディア，Vol.48，SIG16，pp.124（ 2007）

4) 馬場：“講座マルチメディア検索の最先端　第1回　マルチメディア検索の技術動向，”映情学誌，Vol.64，No.1，pp.5863（2010）

5) J. Sivic and A. Zisserman：“Video Google：A Text Retrieval Approach to Object Matching in Videos，”Proc. IEEE International Conference on Computer Vision，Vol.2，pp.14701477（2003）

6) 河合，住吉，柴田，馬場口：“テクスチャ特徴に基づくテレビ番組映像からの高次特徴抽出，”信学技報，PRMU　200889，pp.712（2008）

7) 神谷，高橋，井出，村瀬：“一般物体認識のためのマルチモーダル星座モデル，”信学誌，Vol.J92D，No.8，pp.11041114（2009）

8) Y. Jiang，J. Yang，C. Ngo，and A.G. Hauptmann：“Representations of Keypoint-Based Semantic Concept Detection：A Comprehensive Study，”IEEE Tans. Multimedia，Vol.12，No.1，pp.4253（2010）

9) D. Lowe：“Object recognition from local scale-invariant features，”Proc. IEEE International Conference on Computer Vision，pp.11501157（1999）

10) H. Bay，A. Ess，T. Tuytelaars，and L. V. Gool：“Speed-Up Robust Features（SURF），”Computer Vision and Image Understanding，110，（2008）

11) 勞，山口：“実利用が進む顔画像処理とその応用事例　前編　顔画像処理技術の動向，”情報処理，Vol.50，No.4，pp.319326（2009）

12) 勞，山口：“実利用が進む顔画像処理とその応用事例　後編　顔画像処理の応用事例，”情報処理，Vol.50，No.5，pp.436443（2009）

13) P. Viola and M. Jones：“Rapid Object Detection using a Boosted Cascade of Simple Features，”Proc. IEEE Conf. on Computer Vision and Pattern Recognition，pp.511-518（2001）

14) 松井，後藤，木村，中田，松本，クリピングデル，藤井，八木：“GibbsBoost顔検出と映像監視業務への応用，”映情学誌，Vol.62，No.3，pp.408-413（2008）

15) S. Clippingdale，M. Fujii，and M. Shibata：“Multimedia Databases for Video Indexing：Toward Automatic Face Image Registration，”Proc. IEEE International Symposium on Multimedia，pp.639644（2009）

16) N. Kumar，A. C. Berg，P. N. Belhumer，and S. K. Nayar：“Attribute and Simile Classifiers for Face Verification，”Proc. IEEE International Conference on Computer Vision，pp.365372（2009）

17) S. Satoh，Y. Nakamura and T. Kanade：“Name-it：Naming and Detecting Faces in News Videos，”IEEE Multimedia，pp.22-35（1999）

18) D. Ramanan，S. Baker，and S. Kakade：“Leveraging archival video for building face datasets，”Proc. IEEE International Conference on Computer Vision，pp.1－8（2007）

19) M. Everingham，J. Sivic and A. Zisserman：“‘Hello！My name is ... Buffy’－Automatic Naming of Characters in TV Video，”Proc. British Machine Vision Conference，pp.889－908（2006）

20）Y. Zhang, C Xu, H. Lu, and Y. Huang, “Character Identification in Feature-Length Films Using Global Face-Name Matching,” IEEE Trans. Vol.11, No.7, pp.1276-1288（2009）

21）P. T. Pham, M. Moens, and T. Tuytelaars：“Cross-Media Alignment of Names and Faces,” IEEE Trans. Multimedia, Vol.12, No.1, pp.13-27（2010）

22）N. Dalal and B. Triggs：“Histograms of Oriented Gradients for Human Detection,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.886-893（2005）

23）V. Ferrari, M. Marin-Jimenez, and A. Zisserman：“Pose Search：retrieving people using their pose,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.1-8（2009）

24）I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld：“Learning realistic human actions from movies” Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.1-8（2008）

25）O. Duchenne, I .Laptev, J. Sivic, F. Bach and J.Ponce：“Automatic Annotation of Human Actions in Video,” Proc. IEEE International Conference on Computer Vision, pp.1491-1498（2009）

26）N. Ikizler-Cinbis, R. G. Cinbis, and S. Sclaroff：“Learning Actions From the Web,” Proc. IEEE International Conference on Computer Vision, pp.995-1002（2009）

27）J. C. Niebles, H. Wang, and L. Fei-Fei：“Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words,” J.Comp. Vision, Vol. 79, No.3, pp.299-318（2008）

28）M. Alex and O. Vasilescu：“Human Motion Signatures：Analysis, Synthesis, Recognition,” Proc. International Conference on Pattern Recognition, Vol.3, pp.30456-30460（2002）

29）木谷，岡部，佐藤，杉本：“視覚的文脈を用いた人物動作のカテゴリー学習,”信学論，Vol.J92-D, No.8, pp.1144-1152（2009）

30）渡邉，岡真，鹿毛，鷲見：“コンテクストを用いた動画像からの対象認識の高精度化,”信学論，Vol.J92-D, No.4, pp.521-530（2009）

31）Y. Jiang, J. Wang, S. Chang, and C. Ngo：“Domain Adaptive Semantic Diffusion for Large Scale Context-Based Video Annotation,” Proc. IEEE International Conference on Computer Vision, pp.1420-1427（2009）

32）M. Takahashi, Y. Kawai, M. Fujii, and M. Shibata：“NHK STRL at TRECVID 2009：Surveilance Event Detection and High-Level Feature Extraction,” Proc. TECVID 2009 Workshop（2009）

33）C. Liu, Q. Huang, S. Jiang, L. Xing, Q. Ye, and W. Gao：“A framework for flexible summaraization of racquet sports video using multiple modalities,” Proc. Computer Vision and Image Understanding, 113, pp.415-424（2009）

34）C. Xu, Y. Zhang, G. Zhu, Y.Rui, H. Lu, and Q. Huang：“Using Webcast Text for Semantic Event Detection in Broadcast Sports Video,” IEEE Trans. Multimedia, Vol.10, No.7, pp.1342-1355（2008）

35）高橋，今，長谷山：“アクティブネットを用いたサッカー映像におけるパス可能領域の推定,”信学論（D），Vol. J92-D, No.4, pp.501-510（2009）

36）A. Gupta, P. Srinivasan, H. Shi, and L. Davis：“Understanding Videos, Constructing Plots Learning a Visually Grounded Storyline Model from Annotated Videos,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.2012-2019（2009）

37）M. Takahashi, M. Fujii and N. Yagi：“Automatic pitch type recognition from baseball broadcast videos,” Proc. IEEE International Symposium on Multimedia, pp.15－22（2008）

38）三須，苗村，藤井，八木：“選手フォーメーション解析に基づくサッカーイベント判別法,”映情学誌，Vol.61，No.9，pp.13671375（2007）

39）河合，住吉，柴田，八木，馬場口：“番組紹介テキストの特徴に基づく番組紹介スポットの自動生成,”映情学誌，Vol.64，No.1，pp.85-93（2010）

40）望月，蓼沼，藤井，伊藤：“データベース中の代表的なテクスチャと色を用いて作成した問合せ画像による画像検索,”信学論，Vol.J88-D-II，No8，pp.17291739（2005）

41）J. P. Collomosse, G. McNeill and Y. Qian：“StoryBoard sketches for content based video retrieval”, Proc. IEEE International Conference on Computer Vision, pp.245－252（2009）

42）T. Chen, M. Cheng, P. Tan, A. Shamir, and S. Hu：“PhotoSketch：Internet Image Montage,”ACM SIGGRAPH Asia 2009, No.124（2009）

43）今野，二反田，長谷山：“メロディーとリズムに着眼した音楽信号の類似度に関する一考察,”信学技報，Vol.106，No.534，pp.125128（2007）

44）武，瀧本，佐藤，安達：“特徴点軌跡の不均一性パターンに基づいて同一場面映像検出,”信学論，Vol.J92D，No.8，pp.11531165（2009）

45）P. Asirelli, S. Little, M. Martinelli, and O. Salvertti：“MultiMedia metadata management：A proposal for an infrastructure”,Proc. Semantic Web Applications and Perspectives（SWAP）－3rd Italian Semantic Web Workshop Scuola Normale Superiore（2006）

46）K. Petridis, D. Anastasopoulos, C. Saathoff, N. Timmermann, I. Kompatsiaris, and S. Staab：“M-OntoMat-Annotizer：Image annotation. Linking ontologies and multimedia low－level features,” Proc. Engineered Applications of Semantic Web Session（SWEA）at KES 2006（2006）

47）A. Messina, L. Boch, G. Dimino, W. Bailer, P. Schallauer, W. Allasia, M. Groppo, M. Vigliante, and R. Basili“:Creating rich metadata in the TV broadcast archives environment：The Prestospace project,” Proc. 2nd Int. Conference on Automated Production of Cross Media Content For MultiChannel Distribution（2006）

48）J. Loffler, K. Biatov, C. Eckes, and J. Kohler：“iFinder：An MPEG7based retrieval system for distributed multimedia content,” Proc. ACM Multimedia2002（2002）

49）MPFホームページ，
http://www.nhk.or.jp/strl/mpf/index.htm

50）三浦，山田，住吉，八木，奥村，徳永：“放送番組を素材としたマルチメディア百科事典の自動構築,” 映情学誌，Vol.62，No.1，pp.110-116（2008）

ふじいまひと
藤井真人

1983年入局。札幌放送局を経て，現在，放送技術研究所人間・情報科学研究部主任研究員。この間，CMUに半年間滞在およびATR人間情報通信研究所に出向。神経回路モデル，視覚情報処理，画像認識，映像検索などの研究開発に従事。博士（情報科学）。

しばたまさひろ
柴田正啓

1981年入局 。新潟放送局，放送技術研究所，技術局，放送技術局を経て，現在，放送技術研究所人間・情報科学研究部部長。情報検索，画像データベース，映像ハンドリング技術，番組制作システムなどの研究開発に従事。博士（情報学）。

解説

コンテンツを自動的に推薦するテレビ

住吉英樹　佐野雅規　後藤　淳　望月貴裕　宮崎　勝
藤井真人　柴田正啓　八木伸行

ビデオオンデマンドサービスなど，大量の映像へのアクセスが容易になっている今日，新しいテレビの見方を実現することを目指して開発したコンテンツ推薦テレビの概要と拡張性を持たせたシステムモデルについて紹介する。コンテンツ推薦テレビでは，視聴中のコンテンツに関連したコンテンツをメタデータを利用して自動で検索し，推薦する。複雑な検索操作は不要で，テレビ視聴者の多様な好奇心を満たし，興味を拡大するコンテンツを推薦することができる。

1．まえがき

放送と通信の融合を機に，テレビの役割も変化しつつある。特に，近年の蓄積メディア，インターネット技術の急速な発展は大量の映像の蓄積やネットワークによる配信を可能とした。しかし，映像が大量になるほど見たいものを探し出す作業は困難になる。蓄積された大量のコンテンツを有効に利用するためには，検索機能が重要であるが，テレビの視聴者にキーワード入力などの煩雑な検索操作を要求することは難しい。

当所では，新しい視聴スタイルとして，視聴中のコンテンツに関連したコンテンツを検索し，推薦することでテレビ視聴者の多様な好奇心を満たすことのできるコンテンツ推薦テレビを提案している[1)]。利用者が選択したコンテンツは，その時点での利用者の興味を反映していると考え，視聴中のコンテンツに付与されたメタデータを検索キーとして用い，利用者の検索操作を軽減する。内容に関連したコンテンツを推薦することができるので，視聴者の知的好奇心を満足できると期待している。

本稿では，コンテンツ推薦テレビの機能やサービス要件に基づき，テレビだけではなくPC（パーソナルコンピューター）などにも展開でき，柔軟で汎用性・拡張性のあるコンテンツ推薦システムモデルについて述べる。また，このモデルに基づいて試作したシステムによる基礎的な実験についても述べる。

2．コンテンツ推薦技術の動向

他の利用者の操作履歴や評価を利用した協調フィルタリングによる推薦システムとしてはGroupLens[2)]やMovieLens[3)]などがよく知られており，現在では，インターネット上の多くの通信販売サイト（AmazonやiTunesのGeniusなど）で利用されている。このような推薦システムに用いられている推薦技術はアクセスの少ないコンテンツ（商品）へ誘導するための仕組みとして一般的な技術となりつつある。DVDレンタル会社のNetflix

社がレコメンデーションエンジンの開発コンテストを行う[4]など，映像コンテンツの推薦にも有効性が認められている。しかし，コンテンツへのアクセスに十分な被覆率（カバレッジ）が期待できない利用環境では，対象にならないコンテンツの数が多くなるなど，協調フィルタリングによる推薦は必ずしも有効ではない。また，内容と無関係に，少数の行動や評価によって推薦が行われることがあり，予想外の商品が推薦される場合もある。意外性のある推薦と受け入れられる場合もあるが，利用者の納得を得られないこともある。

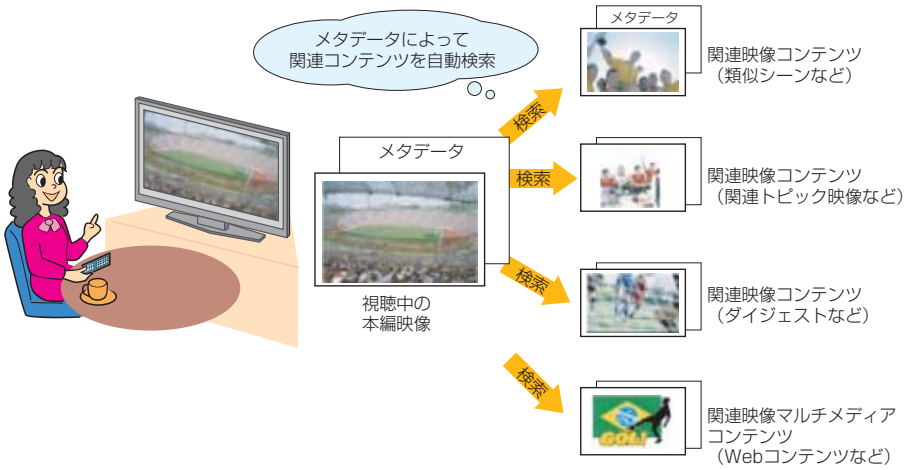
これに対して，対象に関する知識をデータベース化し，その内容を基に推薦を行うコンテンツベースフィルタリングと呼ばれる手法がある。論文やネットニュースの記事内容を対象にした研究が行われている[5][6]。コンテンツベースフィルタリングでは，データベース化されたすべての商品が推薦対象になり，初めての利用者への推薦も可能である。また，商品間の関連を内容情報に基づいて推薦するので，納得されやすい商品を推薦することができる。しかし，コンテンツ内容のデータベースのメンテナンスにはコストがかかり，意外性のある推薦が行われにくいと言われている。

現在では，コンテンツベースフィルタリングと協調フィルタリングを組み合わせたハイブリッド手法の開発が進んでいる[7][8]。推薦結果の評価も推薦精度だけでなく，意外性や推薦に対する納得性など，利用者の満足度に注目した評価法が提案されている[9][10]。

3．コンテンツ推薦テレビ

当所では，視聴中の番組（コンテンツ）に関連するコンテンツに付与されたメタデータを使ってコンテンツを自動的に検索し，推薦する新しいテレビの視聴スタイルを提案している。提案しているシステムでは，コンテンツ提供者が用意する固定的な関連情報だけではなく，種々の検索技術を利用して，アクセス可能な多くの関連コンテンツを推薦することができる。

検索技術を用いるとしても，テレビ番組の視聴を主目的とする受動的な利用者には，PCを使うときのようなキーワードの入力や絞り込みといった複雑な検索操作はなじまない。コンテンツ推薦テレビでは，ユーザーが視聴しているコンテンツがその時点でのユーザーの関心を反映していると考え，視聴中のコンテンツ自体を検索キーとして用いる（1図）。視聴中のコンテンツに付与されたメタデータを利用して関連するコンテンツ



1図　コンテンツ推薦テレビのイメージ

を検索し，推薦・提示を行う。この手法では，検索時の複雑な操作は不要で，視聴者に興味のある領域を「奥行き」と「広がり」を持って自然に拡大していくことができると考えている（2図）。

コンテンツ推薦テレビでは，視聴者の興味を拡大するための推薦を主目的としているので，コンテンツベースフィルタリングの手法を採用した。この手法では，内容に関するメタデータをデータベースとして用い，すべてのコンテンツを推薦対象として内容の関連性によってコンテンツを推薦する。また，利用者が納得できる推薦理由を提示できるというメリットもある。

4．コンテンツ推薦テレビの構成と機能

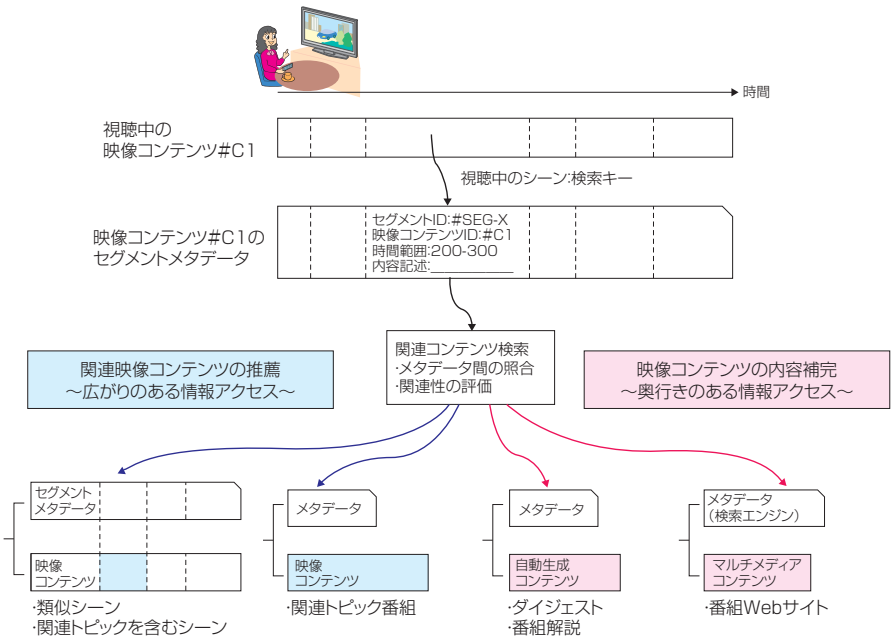
コンテンツ推薦テレビのシステムモデルと構成要素間のインターフェースの考え方と実現できる特徴的な機能について述べる。

4.1　システムモデル

コンテンツ推薦テレビでは，コンテンツの内容に基づく推薦手法をさまざまな用途に適用するので，機能の拡張が容易でなければならない。そこで，以下の4つの要素でシステムモデルを構成する。

（1）表示システム（関連コンテンツの表示）
（2）検索システム（関連性をたどる検索）
（3）メタデータサーバー（メタデータ付与管理）
（4）コンテンツサーバー（コンテンツの配信）

分割した4つの要素を連携して動作させるために，各要素の機能とそれを実現するサブシステム間のインターフェースについて基本的な仕様を定めた。情報の大まかな流れと各部の動作は3図のとおりである。なお，コンテンツの配信については，映像の形式など，実装によりさまざまな方法を取り得るので，コンテンツ推薦テレビでは，コンテンツの実体の場所をメタデータで示すという論理的な機能定義にとどめている。



2図　コンテンツ推薦テレビによる検索の仕組み

3図の構成により，TVやPCなど，多様な表示システムへの対応が可能になる。また，スタンドアローンの小規模なシステムから，複数のサーバー群で構成される大規模なシステムまで，スケーラビリティーを持たせることができる。

システムの構成要素の基本的な動作を3図を用いて説明する。

（1）表示システム（関連コンテンツの表示）

視聴中のコンテンツを一意的に識別する識別子（コンテンツID）と視聴位置（コンテンツ開始点からの再生時間）を検索システムへ送り（①），検索結果として関連コンテンツの内容情報（タイトルや格納位置など）を受け取り（④），内容情報と必要に応じてコンテンツサーバーにリクエストを行い（⑤），関連するコンテンツ（番組の映像など）を表示する（⑥）。

（2）検索システム（関連性をたどる検索）

視聴中のコンテンツIDと視聴位置を表示システムから受け取り（①），視聴中のコンテンツに付与されたメタデータをメタデータサーバーから取得する（②，③）。次に，取得した視聴中のコンテンツのメタデータからクエリー*1を生成し，検索システムにある関連性評価プログラムを用いて関連コンテンツを選定し，関連コンテンツの内容情報を表示システムに返す（④）。

*1 検索に使用する問い合わせ文。

（3）メタデータサーバー（メタデータ付与管理）

検索システムからのリクエスト（②）に応じて，該当するコンテンツのメタデータを検索し，検索システムに返す（③）。

（4）コンテンツサーバー（コンテンツの配信）

表示システムからのリクエスト（⑤）に応じて，必要なコンテンツを配信する（⑥）。ただし，コンテンツの格納場所やアクセス方法はメタデータ内に記述されているとする。

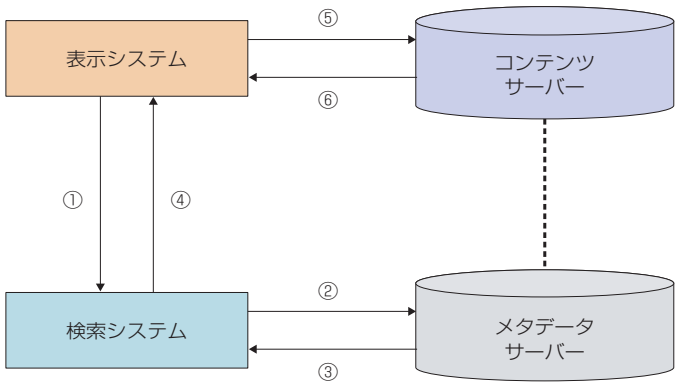
4.2 システム間インターフェース

（1）検索インターフェース

表示システムの発行するリクエストは以下の情報を持つ。

・コンテンツID
・視聴位置
・返却要求順位，返却数

リクエストに含まれる情報はシンプルであり，さまざまな用途に対応でき，表示システムへの実装も容易である。



3図　コンテンツ推薦テレビのシステムモデル

検索システムは検索処理を行い，メタデータサーバーから多数のコンテンツ情報を取得し，以下の情報を１つのコンテンツ情報として表示システムに返す。

・コンテンツID
・関連ラベル
・類似度
・コンテンツ実体URL
・タイトル，サブタイトル
・コンテンツ（番組）概要
・ジャンル，ジャンルコード
・放送日，放送時間

（2）メタデータ取得インターフェース

コンテンツIDをクエリーとして，該当するコンテンツのメタデータを取得するインターフェースである。当所では，メタデータ制作にかかわるフレームワークをメタデータ制作フレームワーク（MPF：Metadata Production Framework）＊2として提案している[11]。メタデータの検索には，MPFで定義した検索インターフェース＊3を用いることを基本としているが，コンテンツ推薦テレビへ実装する際には，検索システムとメタデータを一体として実装する形態も考えられるので，検索クエリーのキーとしてコンテンツIDを用いること以外に受け渡す情報や戻り値の形式などは定義していない。

＊2
本特集号の解説「メタデータ制作フレームワーク」参照。

＊3
MPFでは蓄積したメタデータの操作（編集，検索など）を行う関数群をインターフェースとして定義している。

4.3　実現できる推薦機能

システムモデル（4.1節）とインターフェース（4.2節）の定義により，自由度の高い検索・推薦機能が実現できる。

（1）コンテンツの内容に応じた検索

視聴中のコンテンツに付与されたジャンルなどの属性情報に応じて，検索システムに組み込まれた複数の関連性／類似度評価機能（検索処理）から適切なものを選択し，視聴中のコンテンツの内容に適した関連コンテンツを利用者に推薦する。複数の検索処理を切り替えることで，内容に適した検索を行うことが可能になる。表示システム側では検索処理の種類や場所の指定は不要であり，表示システムを実装する際の負荷や実行時の負荷を低減している。検索システムの自由度は高く，検索機能の拡張も容易である。

（2）コンテンツの粒度＊4に応じた検索

コンテンツIDと視聴位置をキーとして，コンテンツに付与されたメタデータを視聴位置ごとに取得する。検索のクエリーは視聴位置ごとに変化するので，ニュースの項目やシーン（項目の一部分）ごとに最適な検索結果を得ることができる。メタデータの粒度と検索処理を組み合わせて，コンテンツの粒度に対応した３種類の関連コンテンツを検索することができる。

・コンテンツからコンテンツ
・シーンからコンテンツ
・シーンからシーン

＊4
番組単位，シーン単位，ニュース項目単位など，番組内の意味的な大きさの違い。

５．システムの試作とコンテンツの推薦動作

コンテンツ推薦テレビの具体的なイメージをわかりやすく説明するために試作したシステムについて述べる。試作したシステムでは，アーカイブやVOD（Video On Demand）などに蓄積されたコンテンツをネットワークに接続されたTV受信機あるいはWebブラウザーで視聴するという形態を想定しており，４図のような構成である。検索

システムのインターフェース（4.2節）はHTTP SOAP（Simple Object Access Protocol）Webサービス*5として実装している。

以下，コンテンツ推薦テレビの中核をなす検索システムと表示システムの概要を説明する。

5.1 検索システム

検索システムには3種類の検索機能を実装している。視聴中のコンテンツのジャンルとメタデータに基づいて，どの機能を使用するのかを選択する。

（1）言語処理により類似した話題をたどる手法

電子番組表の概要情報など，コンテンツ内容を記述した関連テキスト（字幕，音声認識による発話の書き起こし，番組ホームページ等）を言語処理で解析する。文章中の特徴的な単語の出現頻度に着目して文章間の類似度を計算し，話題の類似したコンテンツを検索する[12)]。ニュースでは音声認識を利用してテキストに変換し，項目ごとに関連コンテンツを抽出する。また，類似度に大きく寄与する単語（人名や話題の内容）に着目し，どのような関連があるのかを推定する手法を実装している。

（2）言語処理により同一イベントをたどる手法

野球のニュース項目のアナウンスコメントを言語処理で解析し，ニュースで取り上げたイベントごとに試合進行や選手名を抽出する。抽出したデータと専門業者が手動で入力した投球ごとの試合データとを比較して，VODサーバー等に蓄積された中継映像から同一イベントのシーンを検索する手法を実装している[13)]。

（3）映像処理により同一シーンをたどる手法

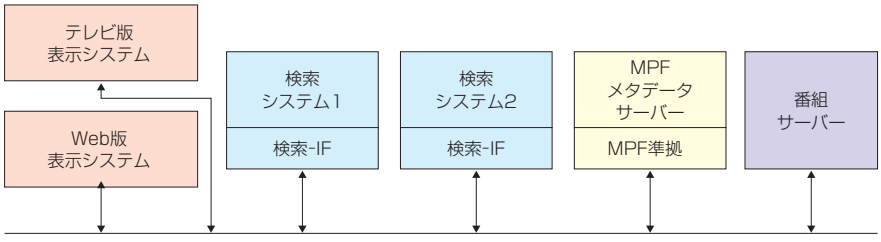
野球のニュース映像を映像処理で解析し，サーバー等に蓄積された中継映像からニュースのシーンと同じシーンを検索する手法を実装している[14)]。この手法では，12分割した画像領域から平均色（RGB）とエッジ画素数の4種類の量を計算して，合計48次元（4種類×12領域）のベクトルを抽出する。このベクトルを一定のフレームにわたって比較して，ニュース映像に最も類似したシーンを中継映像から抽出する。

5.2 表示システム

利用環境やポインティングデバイスの有無など，表示形態や操作性が変更できるシステムモデルであるかどうかを確認するために，TV受信機を想定した表示システムとWebブラウザーを想定した表示システムを構築し，動作を検証した。以下，言語処理により類似した話題をたどる手法を用いて検索を行い，関連するコンテンツを推薦表示する方法について述べる。

（1）TV型の表示システム

TV型の表示システムでは，テレビ視聴という受動的な利用状況を考慮し，リモコンでの操作を仮定した。全画面のコンテンツ表示状態（標準の視聴状態）から，コンテンツ



4図　試作したシステムの構成

＊5
インターネットで使用されるHTTP通信プロトコルを使用して，他のコンピューターのプログラムを呼び出すための通信規約。

推薦モードへの移行ボタンを押すと，５図のような表示になる。視聴中のコンテンツを左上に縮小表示するとともに，画面右側に３つの推薦コンテンツを推薦理由の関連キーワードと共に表示する。推薦コンテンツの数については，リモコンでの操作性を考慮して，暫定的に３つとしているが，最適な数については，今後の課題である。また，表示数が限定されるので，連続ドラマなど，検索元と同一の番組名を持つシリーズ番組は除外している。

TV型の表示システムでは，テレビ視聴の特質を考慮し，推薦コンテンツをすぐに視聴するような遷移をせずに，興味を引いたコンテンツをマークするブックマーク機能を付加している。ブックマークしたコンテンツは，ブックマーク一覧表示機能から，後で視聴できるようにしている。

（２）Web型の表示システム

Webブラウザー上で表示可能なFlash*６を用いた表示システムである。TV型の表示システムとは異なり，マウスなどのポインティングデバイスが利用できること，画面と利用者の距離が近いことなど，PCでの利用形態を考慮して，６図に示すように数十個の推薦コンテンツを一覧にして表示する。なお，画面の中心に表示している映像は視聴中の映像である。

検索システムからのコンテンツ情報には，類似度と関連の種類を示す語が含まれているので，同一の関連語を持つコンテンツを１つのクラスターとしてまとめ，類似度の総和の高いものから３クラスターと，残りのすべてのクラスターを１つにまとめた計４クラスターを画面の中央付近から４隅の方向に向かって表示する。類似度の高いコンテンツを視聴中の映像の４隅の近くに配置し，表示するサムネイルのサイズ（５段階）を大きくし，視聴中のコンテンツと大きな関連があることを示している。画面の４隅には，そのコンテンツとどのような関連があるのかを示す単語を表示している。協調フィルタリングを用いた推薦に関する研究では，評価の分布など，推薦の根拠や度合いを提示することが利用者の安心につながるという報告[9]があり，類似度に大きく寄与する単語（人名や話題の内容）を示すことで，利用者の納得性を高めている。

推薦コンテンツとして表示されたサムネイルにマウスカーソルを重ねると，コンテンツ情報が吹き出しの中に文字で表示され，コンテンツの冒頭の映像が流れる。興味のあるコンテンツであれば，サムネイルをクリックしてコンテンツを選択する。選択されたコンテンツは中央のコンテンツ視聴領域に表示されるとともに，そのコンテンツに関連

＊６
音声，画像，動画などを組み合わせて動きのあるWebコンテンツを作成するためのソフトおよび作成されたコンテンツ。Adobe Systems社が開発。

５図　TVを用いた表示システムの画面例

するコンテンツが再度検索されて周りに配置される。このように，利用者はあたかもWebサーフィンをするように，興味のある映像コンテンツを次々に視聴していくことができる。

5.3　試作システムによるコンテンツの推薦

5.1節で述べた3種類の検索を行うために，約4,000番組の番組概要文を中心とするメタデータと，数本の野球中継番組の映像特徴量（ベクトルデータ）をメタデータサーバーに蓄積した。TV型の表示システムとWeb型の表示システムは同一の検索システムとメタデータを用いてコンテンツを推薦する。なお，ニュース番組では，時間的に区切られた区間に，適切なセグメントメタデータを付与することで，動的に変化するコンテンツに逐次適応したコンテンツを推薦することができる。言語処理や画像処理といった異なる検索処理を同じ枠組みに実装しているので，拡張性の高いシステムモデルと言える。

試作システムでは、ジャンルによって検索処理を切り替えるが，具体的な内容に依存した検索処理をする必要があること，検索の要求タイミングを適切に行う必要があることなどの課題がある。

6．まとめ

新しいテレビの視聴スタイルであるコンテンツ推薦テレビについて述べた。紹介したシステムモデルとインターフェースは複数の表示システムや検索処理へ対応可能な拡張性を持っている。

コンテンツの推薦で重要なことは，関連性を求めるための検索処理であり，どのような関連があると判断するかということである。試作したシステムでは，類似度を指標としたが，利用者の満足度を指標とすれば，類似度による深さ方向の関連だけでなく，意外性のある広がり方向に関連のあるコンテンツも提示することができるようになる。また，操作性や推薦に対するユーザーの満足度についての検証・評価は，今後の課題である。また，さまざまな関連性を抽出する検索技術やメタデータ抽出など，メディア処理

6図　Webブラウザーを用いた表示システムの画面例

すみよしひでき
住吉英樹

1980年入局。広島放送局を経て，1984年から放送技術研究所にて，コンピューターを応用した番組制作システム，メタデータ制作システムの研究に従事。現在，放送技術研究所人間・情報科学研究部専任研究員。博士（工学）。

さ　の　まさのり
佐野雅規

1994年入局。仙台放送局を経て，1997年から放送技術研究所にて，コンテンツ制作，メタデータ制作技術，メディア情報処理などの研究開発，ARIB,MPEG,EBUなどの標準化活動に従事。現在，放送技術研究所人間・情報科学研究部主任研究員。博士（情報学）。

技術の研究を進めるとともに，大規模アーカイブスなどの映像を実用的な速度で検索・推薦するためのインデックスのあり方も含めて高速化の検討が必要である。今後，多数の検索システムをニーズに合わせて適切に選択する仕組みや，個人ごとに適切なコンテンツを推薦するための方法などについても検討を進めていく。

7．謝辞

TV版のコンテンツ推薦テレビ表示システムはパナソニック（株）と共同で開発した。深く感謝する。

後藤　淳（ごとう　じゅん）

1993年入局。高松放送局を経て，1998年から放送技術研究所にて，知的インターフェース，自然言語処理の研究に従事。現在，放送技術研究所人間・情報科学研究部専任研究員。

望月貴裕（もちづきたかひろ）

1996年入局。放送技術局報道技術センター中継制作部を経て，1998年から放送技術研究所にて，画像および映像解析の研究に従事。現在，放送技術研究所人間・情報科学研究部専任研究員。博士（工学）。

宮崎　勝（みやざき　まさる）

1997年入局。名古屋放送局を経て，2000年から放送技術研究所にて知識処理（主にエージェント技術，オントロジー技術）の研究に従事。現在，放送技術研究所人間・情報科学研究部専任研究員。

参考文献

1） 藤井，柴田，住吉，後藤，佐野，望月，宮崎，八木：“CurioView：検索技術を活用した新しい視聴スタイルの提案，”映情学年大，6−5（2009）
2） P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J.Riedl：“GroupLens：An Open Architecture for Collaborative Filtering of Netnews，”Proc. Conf. on Computer Supported Cooperative Work, pp.175−186（1994）
3） MovieLens , http://www.movielens.org/
4） Netflix Prize, http://www.netflixprize.com/
5） K. Lang：“NewsWeeder：Learning to Filter Netnews，”Proc. of the 12th International Conf. on Machine Learning, pp.331−339（1995）
6） S. Lawence, C. Lee Giles and K. Bollacker：“Digital Libraries and Autonomous Citation Indexing，”IEEE Computer, Vol.32, No. 6, pp.67−71（1999）
7） A. Das, M. Datar and A. Garg：“Google News Personalization：Scalable Online Collaborative Filtering，”Proc. 16th International Conf. on World Wide Web, pp.271−280（2007）
8） K. Yu, A. Schwaighofer, V. Tresp, W. Ma and H. Zhang：“Collaborative Ensemble Learning：Combining Collaborative and Content−Based Information Filtering via Hierarchical Bayes，”Proc. Uncertainty in Artificial Intelligence, Vol.19, pp.616−623（2003）
9） R. Sinha and K. Swearingen：“The Role of Transparency in Recommender Systems，”Proc. SIGCHI Conf. on Human Factors in Computer Systems, pp.830−831（2002）
10） K. Sweringen and R. Shinha：“Beyond Algorithms：An HCI perspective on recommender systems，”ACM SIGIR Workshop on Recommender Systems（2001）
11） M. Sano, H. Sumiyoshi, M. Shibata and N. Yagi：“Metadata Production Frameowrk（MPF）Version 2.0，”Proc. ACM Multimedia 2009, pp.1017−1018（2009）
12） J. Goto, H. Sumiyoshi, M. Miyazaki, H. Tanaka, M. Shibata, A. Aizawa：“Relevant TV Program Retrieval using Broadcast Summaries，”Proc. 14th ACM International Conf. on Intelligent User Interfaces, ACM, pp.411−412（2010）
13） 宮崎，住吉，後藤，藤井，柴田：“スポーツニュースの言語情報を利用したプロ野球映像推薦システムの試作，”情報科学技術フォーラム講演論文集，6F−4, FIT 2009 Sep（2009）
14） 望月，藤井，八木，篠田：“投球の次ショットに重きを置いたシーンのパターン化と離散隠れマルコフモデルを用いた野球放送映像の自動イベント分類，”映情学誌，Vol.61, No.8, pp.1139−1149（2007）

藤井真人（ふじいまひと）

1983年入局．札幌放送局を経て，現在，放送技術研究所人間・情報科学研究部主任研究員。この間，CMUに半年間滞在，およびATR人間情報通信研究所に出向。神経回路モデル，視覚情報処理，画像認識，映像検索などの研究開発に従事。博士（情報科学）。

柴田正啓（しばたまさひろ）

1981年入局．新潟放送局，放送技術研究所，技術局，放送技術局を経て，現在，放送技術研究所人間・情報科学研究部部長．情報検索，画像データベース，映像ハンドリング技術，番組制作システムなどの研究開発に従事。博士（情報学）。

八木伸行（やぎのぶゆき）

1980年入局。甲府放送局，放送技術研究所，技術局，編成局を経て，現在，放送技術研究所研究企画部部長。画像・映像・メディア情報処理，コンピューターアーキテクチャー，コンテンツ制作技術，デジタル放送などの研究開発に従事。博士（工学）。

解説

メタデータ制作フレームワーク

佐野雅規　住吉英樹　藤井真人　柴田正啓　八木伸行

所望の映像コンテンツをより速くより正確に取得したいという要望を満たすためには，映像コンテンツに検索のためのメタデータ（内容記述情報）を付与する必要がある。特に，あるシーンを見つけたい場合には，映像の時間軸に沿った意味内容を記述するメタデータが必要である。現在，このメタデータを付与する作業は人手に頼らざるを得ない状況であるが，当所では，メディア解析技術を組み合せて，このようなメタデータをできるだけ効率的に生成するための環境として，メタデータ制作フレームワーク（MPF：Metadata Production Framework）を提案している。本稿では，メタデータ制作フレームワークの概要を紹介する。

1．まえがき

昨今の技術進歩により，コンピューターによる映像データの圧縮・蓄積・再生など，映像を物理的に扱うことは比較的容易になってきている。しかし，ある特定のシーンを探すなど，映像を意味的に扱うことについては，いまだに大きな壁がある。特に，放送局など，大量の映像コンテンツを扱う組織では，映像を意味的に扱う機会が多く，技術的な解決が期待されている分野であり，映像コンテンツから内容を反映したメタデータ（内容記述情報）を抽出し，管理することが求められている。

当所では，放送局および映像コンテンツを制作・配信する事業者の立場で，番組など映像コンテンツにメタデータを効率よく付与するためのフレームワークとして，メタデータ制作フレームワーク（MPF）を提案している。このフレームワークでは，目的の映像コンテンツに複数の研究者や機関が連携してメタデータを付与することが可能であり，技術提供者も利用者も双方にメリットがある。2006年にバージョン１を当所のWebページで公開し，その後もさまざまなプロジェクトでの実験と検証を重ね，2008年にはネットワーク対応への拡張を行ったバージョン２を公開し，現在も仕様の改良を続けている。本稿では，最新のMPF仕様の骨子とリファレンスソフトウエアを用いたMPFの基本的な利用シナリオについて紹介する。また，これまでの活動や今後の展望についても述べ，MPFへの理解促進と連携への協力を促したい。

2．メタデータ制作フレームワーク（MPF）

2.1　MPFの位置づけと対象メタデータ

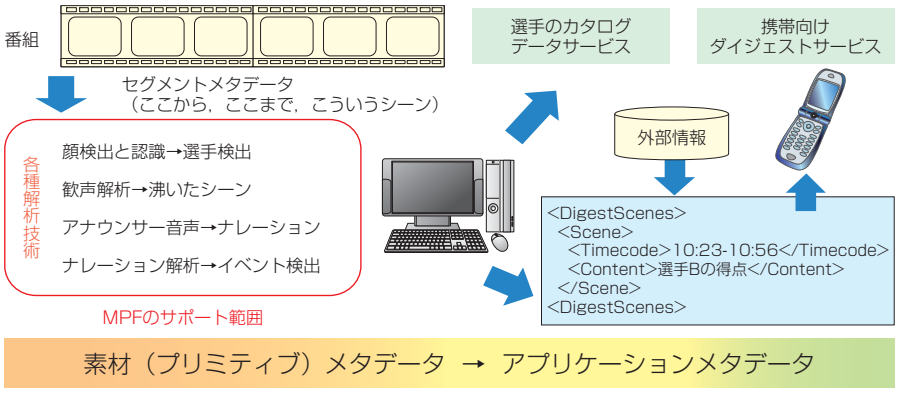
メタデータは「データのためのデータ」と説明されることが多く，コンテンツに関連するさまざまな情報をすべてメタデータと呼ぶこともある。メタデータは，一般的に，

ある特定の用途，特定のアプリケーションで効率よく利用されるので，必要最低限の情報が独自のフォーマットで表現されており，その種類は千差万別である。

放送局においては，番組に対して，検索などさまざまな用途を目的としてメタデータを付与している。1図はMPFを設計するにあたって想定した将来のメタデータの制作フローである。メタデータの制作体系を2段階に分け，第1段階（1図の左側）では共通に利用できるメタデータの作成を，第2段階（1図の右側）ではそれらを基に特定のアプリケーションに特化したメタデータを作成する。第1段階のメタデータを素材メタデータまたはプリミティブメタデータと呼び，第2段階のメタデータをアプリケーションメタデータと呼ぶ。多種多様なアプリケーションメタデータを制作する場合に，それぞれのメタデータを制作するために個別のシステムを開発するのでは膨大なコストがかかる。逆に，1つのシステムとして実現するとシステムが過度に複雑になる。そこで，共通の素材メタデータを作る第1段階と，これを基にアプリケーションメタデータを制作する第2段階に分けて，全体的なコストの削減を図っている。このような制作体系を効率よく稼働させるためには，第1段階の素材メタデータをいかにコストをかけずに精度よく生成するかということが鍵となる。本稿で紹介するMPFは，この素材メタデータを精度よく自動的に生成するための環境を提供することを目的としている。

一般的に，番組に関連するメタデータはその記述する対象範囲の大小によって2種類に分けることができる。1つは番組全体にかかわるタイトルや制作者など書誌情報*1に当たるもので，これをプロダクションメタデータと呼ぶ。通常，人手によって付与される。放送局においては番組管理のために付与されており，データベース（DB）も整っている。他の1つは番組のある時間区間に対して付与されたメタデータである。セグメントメタデータと呼ばれ，その基本構成要素は，ここからここまでという時間区間の境界情報と，その区間の内容や使用にかかわる情報である。セグメントメタデータで付与される情報にはさまざまな種類があり，付与する手法も異なっている。例えば，ある区間映像に対する再使用上の制約や著作権などの情報はこの1つであり，人手によって入力される。また，映像の構図や色合いなどの低次の特徴は映像解析によってある程度自動で抽出することができる。映像中のイベントやシーンの意味などの高次の特徴は検索などをするユーザーに最も有用なメタデータであるが，自動抽出は難しく，ほとんど付与できていないのが現状である。MPFで扱うメタデータは低次から高次の特徴までであり，さまざまなメディア処理を組み合せることで効率よく生成することを目的としている。

＊1
タイトル，著者，出版社など書籍を特定するための情報。



1図　メタデータの制作フロー

2.2　システムモデル

２図はMPFにおけるメタデータ制作システムのモデルである。モジュール群とそれらを制御するためのコントローラーの２種類で構成されるシンプルなモデルである。モジュールは２種類あり，１つは蓄積モジュールと呼ばれている。生成したメタデータを蓄積・管理するDBを持つモジュールで，コントローラーまたは他のモジュールからの要求に応じて，メタデータDBを検索または更新する。他の１つは処理モジュールと呼ばれ，MPFにおいて最も重要なメタデータ制作にかかわる各種処理機能を提供する。各種処理機能には，メタデータの生成・加工・利用・削除など，メタデータのライフサイクルに絡んだすべての操作が含まれている。コントローラーはユーザーとシステムの仲介を行い，目的のメタデータを生成するために必要なモジュール群を制御する。処理モジュールとのやり取りのほかに，生成したメタデータを保管・活用するために，蓄積モジュールとのやり取りを制御する。

MPFではメタデータ制作にかかわるさまざまな処理を統一されたインターフェース（IF）を持つモジュールとして実装しており，それらを組み合わせることで，目的のメタデータを制作するという考え方に基づいている。メタデータの制作を連携して行うための方法として，作成したメタデータだけを交換するのではなく，個々の処理モジュール自体も交換することが可能である。このようなシステムモデルを支障なく機能させるためには，メタデータの表現（メタデータモデル）と，モジュールやコントローラーの間でのデータのやり取り（インターフェース）を規定する必要がある。

３．MPF仕様の規定項目

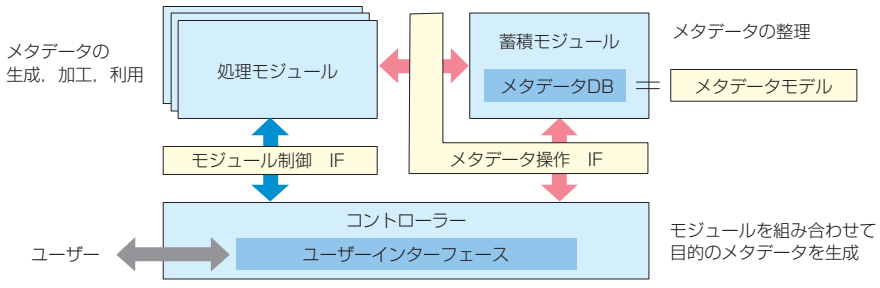
3.1　メタデータモデル

MPFのメタデータモデルには，メタデータの国際標準であるMPEG−7[2)]のサブセットを採用した。MPFのメタデータをXML（Extensible Markup Language）形式*２で表現し，その構造などをXMLスキーマ*３によって定義した。ただし，一部のスキーマで表現できない制約は文章で規定した。MPEG－７からのサブセットの選択方法については，当初は番組を記述するための最小限のものにとどめ，実証実験を進めていく過程で，必要になれば順次拡張するという方針で進めた。そのためMPFバージョン１では，映像や音の低次の特徴は対象外とし，基本的にテキストで記述されたメタデータだけを対象とした。MPFバージョン２では，映像や音の低次の特徴も含め，MPEG－７で規定されていないメタデータについても外部ファイルに保存し，そこへのポインターを保持するという形で拡張した。

３図はMPFのメタデータモデルとシステムモデルの核である処理モジュールの動作を

＊２
異なるコンピューター間でデータを交換するために規定された規格。タグでマークアップすることで，データの意味づけや構造化ができる。

＊３
XMLの構造を規定するための言語。スキーマ自体もXMLを用いて規定されている。



2図　MPFのシステムモデル

示したものである。MPFでは，単一の映像コンテンツ（番組）を対象としており，図の中央がメタデータの構造である。先に述べたプロダクションメタデータは基本情報の中にある。メタデータの作成方法ごとにセグメントブロックを作り，その中にその方法で作成したセグメントメタデータをセグメントユニットとして入れる。１つの番組メタデータの中に，セグメントブロックは幾つでも生成することができ，セグメントブロックの中には必要なだけセグメントユニットを生成することができる。作成単位としてはショット区間*4や発話区間などがある。

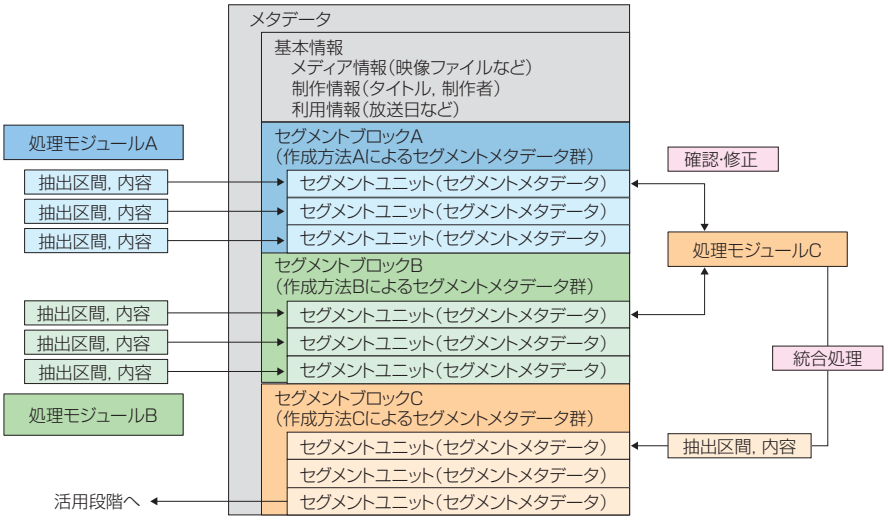
ここで，処理モジュールを用いたセグメントメタデータの生成についてニュース番組を例にして説明する。３図の処理モジュールAは番組の映像を解析して１つ１つのニュース項目を検出し，それをセグメントブロックAにまとめて格納する。同様に，処理モジュールBはアナウンサーの発話を認識し，１文１文をセグメントユニットとしてセグメントブロックBに格納する。処理モジュールCは，これら２つのモジュールの結果を用い，各ニュース項目の中に含まれるアナウンサーの発話内容を言語解析し，抽出した主題などのサマリーを内容情報として付与する。この例では，３つの処理モジュールを組み合せた一連の処理で，各ニュース項目に関するメタデータを生成する。このようにセグメントブロックを単位としてモジュールにより提供される処理を組み合せ，また，他のモジュールの結果を再利用して目的のメタデータを生成する。

3.2　インターフェース

MPFでは，生成したメタデータを送受するためのメタデータ操作インターフェースと，モジュールを制御するためのモジュール制御インターフェースの２種類を規定している。どちらもWebServices*5による実装を基本としているが，モジュール制御インターフェースについては，処理モジュールを容易に開発するためにWindows DLL*6による実装も可能としている。以下，２種類のインターフェースの概要を説明する。

3.2.1　メタデータ操作インターフェース

蓄積モジュールに蓄えられるメタデータを操作するためのインターフェースである。システムへの実装を容易にするために，高レベルと低レベルの２種類のインターフェースを定義しており，すべての蓄積モジュールは高レベルインターフェースの実装を必須としている。高レベルインターフェースはメタデータモデルの中の５つの構造要素（番



3図　MPFのメタデータモデルと処理モジュールの動作の関係

＊4
通常，カメラで撮影を開始して停止するまでの一連の映像区間。映像コンテンツを構造化する際に最も基本となる単位。

＊5
ネットワークを介してWeb上のアプリケーションを利用するための技術。

＊6
Windows上で複数のアプリケーションが共通して利用する機能をまとめてファイルとして保存したプログラム。

組全体，セグメントブロック，セグメントユニット，番組全体の基本情報，番組全体とセグメントブロックのヘッダー情報）を操作単位としている。ネットワークを介した複数のプロセスによって１つのメタデータを更新できるようにするために，書き込み操作権限によるメタデータ操作を用意している。具体的には，更新対象とするメタデータにロックをかけ，他からの更新を禁止した状態にし，更新が終了した後でロックを解除する。これにより複数プロセスの衝突によるデータの破損を防ぐことができる。低レベルインターフェースとしては，XMLで記述されるデータのどの部分にでもピンポイントで自由にアクセス可能なクエリー規格であるＷ３Ｃ（World Wide Web Consortium）*7のXQuery*8を採用している[3]。また，XML要素の更新などについては，現在，Ｗ３Ｃにおいて勧告案であるXQuery Update Facility*9の採用を予定している。

＊7
インターネットのWWW技術に関する標準化団体。

＊8
2007年１月に勧告となったXMLデータに対して問い合わせを行うための言語。

＊9
XMLデータの更新などを要求するための言語。

3.2.2　モジュール制御インターフェース

モジュール制御インターフェースは処理モジュールには必ず実装されており，MPFでは４種類の関数群（モジュールの初期化，モジュールのプロパティーと処理に必要なパラメーターの取得，処理に必要なパラメーターの設定，モジュールの動作制御（開始・停止））を規定するシンプルな構成とした。コントローラーは，これらのインターフェースを介して必要なモジュールを動作させて目的のメタデータを生成する。

４．リファレンスソフトウエア

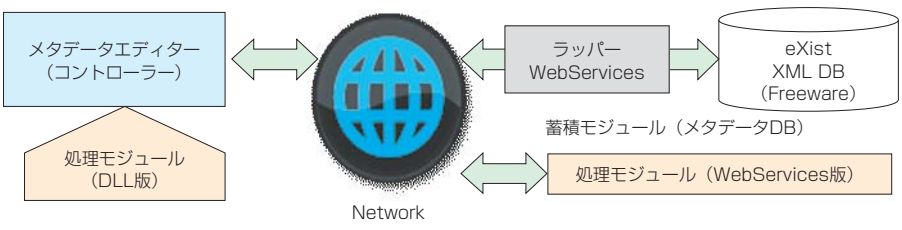
MPFのWebページ[1]ではMPFの仕様書のほかにリファレンスソフトウエアとそのマニュアルを公開している。MPFの基本動作を確認するために無償で配布しており，商用でなければ自由に利用することができる。提供しているソフトウエアは３種類で，コントローラーの機能を持つメタデータエディターと，蓄積モジュールを構築するためのラッパー*10と，サンプルとしての各種処理モジュールである。**４図**はこれらのリファレンスソフトウエアによる実験環境を示している。ソフトウエアは必要なものだけを組み合せて使用することができ，その組み合せによってテストできるMPFの機能が異なる。なお，MPFの仕様書を含め，これらのソフトウエアとそのドキュメントはすべて日本語版と英語版を用意している。

＊10
前処理を行うソフトウエア。

以下，個々のソフトウエアについて簡単に説明する。

4.1　メタデータエディター

MPFシステムモデルにおけるコントローラーの役割を果たし，MPFの基本的な機能を検証することができる。エディター上では，開発したさまざまな処理モジュールを組み合せて動作させることができる。また，人手によるメタデータの修正も可能である。最終的なメタデータは外部ファイルとして保存することもできるし，蓄積モジュールを構築してそこに保存することもできる。**５図**はエディターの操作画面を示している。画面



4図　リファレンスソフトウエアによる実験環境

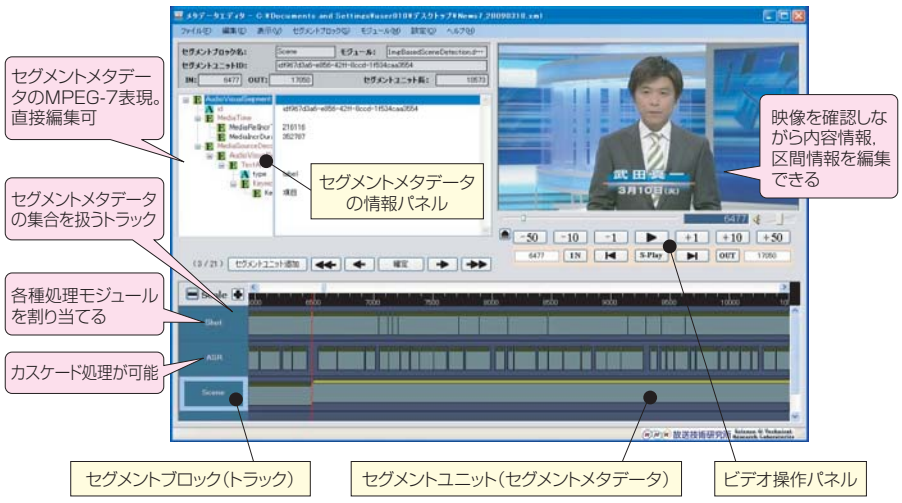
は大きく分けて３部構成となっている。左上が選択されたセグメントユニットの情報表示部，右上が映像コンテンツ表示部，下半分がセグメントブロックとそれに含まれるセグメントユニットの表示部である。処理モジュールはセグメントブロック（トラック）に対して１つ割り当てることが可能で，処理を開始すると抽出されたセグメントユニット（セグメントメタデータに相当）の出力が描画される。トラック上にあるセグメントユニットをクリックするとその内容がエディター画面左上に表示される。選択されたセグメントユニットの内容はMPEG－７に準拠したXMLの木構造で表現されており，直接，編集が可能である。また，エディター上では複数のトラックに同じ処理モジュールを割り当て，それぞれを違ったパラメーターで動作させることも可能であり，パラメーターの違いによる処理結果の違いを視覚的に容易に把握することができる。そのほか，モジュールをカスケード接続する（他のモジュールの出力結果を入力とする）ことも可能であり，正解データをトラックに作成しておけば，それと処理モジュールの結果とを比較して，精度などを数値化して外部ファイルに出力するような評価処理モジュールを開発することもできる。更に，グラフを表示する特別なトラックも実装しており，決められたフォーマットの時刻情報付き数値列を選択した色や形式で描画することもできる。この機能により，メディア解析処理に関連する時系列データを視覚的にわかりやすく確認することができる。

4.2　蓄積モジュールのためのラッパー

MPFシステムモデルにおける蓄積モジュールを構築するためのプログラムである。データベースそのものには，フリーのネーティブXMLデータベース*11であるeXist*12を利用した[4]。ラッパーはeXistのデータベースを扱うメタデータ操作インターフェースを実装しており，WindowsのIIS（Internet Information Server）*13上に構築するWebServicesとなっている。従って，WindowsのOSにeXist，IIS，ラッパーをインストールすることで，ネットワークを介してどこからでも利用可能な蓄積モジュールが構築できる。

4.3　処理モジュール

MPFの処理モジュールとして，画像，音，その他の処理を行うサンプルプログラムをソースコード付きで５種類提供している。これらにはDLL実装とWebServices実装の両



5図　メタデータエディターの操作画面

＊11
XML形式の情報をそのままの形式で扱うことのできるデータベース。

＊12
オープンソースのネーティブXMLデータベースで，ほとんどのOSで動作する。

＊13
マイクロソフトのインターネットサーバーを構築するソフトウエア。

方のソースコードが付いている。更に，モジュール開発者のために，プログラムの骨格だけを持つスケルトンモジュールもソース付きで用意されている。スケルトンモジュールのソースにはプログラムを作成するための指示が付いており，それに沿って自身の研究開発した内容解析処理を埋め込めば，MPF仕様の処理モジュールができ，メタデータエディターからそれを実行してテストすることができる。

5．これまでの活動と今後の予定

MPFは2006年５月に公開した後，国内外のプロジェクトに参加・展示をして，その有用性をアピールしてきた。例えば，2007年度からの情報大航海プロジェクト[5]では，MPFを映像の意味理解のための共通基盤として提案し，多くのメディア処理をモジュール化した。また，自動コンテンツ解析技術（音声認識，顔画像認識，テキスト解析等）に基づくメタデータ付与技術の評価を目的としたEBU（European Broadcasting Union）のP/SCAIE[6]プロジェクトでは，評価のための共通のメタデータフォーマットの基本仕様の１つとして検討されている。今後，これらの活動を続けるとともに，より現実的なシステムとなるように仕様の再検討を行っている。

予定している仕様の拡張は以下のとおりである。

5.1　内容記述対象の空間的分割への拡張

現在の内容記述の対象領域は，映像や音声を時間的に分割したある時間区間であり，映像を構成する画像（フレーム）では全体領域を記述対象としている。最近，映像や画像の空間的な一部分を特定し，そこへメタデータを付与することは多く，MPF仕様の空間的な分割への拡張要望がある。MPEG－７には，この機能が含まれており，それを含む仕様に変更することで，空間的な分割に対応できるようにする予定である。

5.2　リンクメタデータ付与への拡張

リンクメタデータとはコンテンツのある部分と別のある部分との間の関係を記述するメタデータである。単一のコンテンツ内だけではなく，複数のコンテンツ間に付与する場合もある。最近の情報抽出の研究分野では，大量のコンテンツを対象として，このようなさまざまな関係を抽出するということが盛んに行われている。この分野においてもMPFを利用したいという要望があり，仕様の拡張を検討している。現バージョンのMPFは単一のコンテンツだけを対象としており，これを複数対象となるように拡張し，リンクメタデータを効率よく扱うことができるように仕様の変更を検討している。

5.3　リファレンスソフトウエアの改修

5.1節および5.2節で述べたような機能を拡張するためのソフトウエアの改修のほかに，以下の２点の改修を検討している。１つはメタデータエディター上でセグメントメタデータを編集する際のGUI（Graphical User Interface）の改修である。現在のメタデータエディターでは，画面に木構造で表現されたMPEG－７準拠のXMLデータを直接操作することはできるが，そのためにはMPEG－７の知識が必要である。そこで，ユーザーがMPEG－７の知識が無くても利用できるように，この部分のGUIをカスタマイズ可能なプラグインモジュール方式として設計し直すことを考えている。他の１つは現在のメタデータエディターで人手によって行っている一連のモジュールの実行処理をバッチ処理*14で行えるようにすることである。

6．むすび

映像コンテンツを対象にして，内容を記述するメタデータを効率よく付与するための

＊14
一定期間もしくは一定量のデータを蓄積し，一括処理を行う方式。

佐野雅規（さのまさのり）

1994年入局。仙台放送局を経て，1997年から放送技術研究所にて，コンテンツ制作，メタデータ制作技術，メディア情報処理などの研究開発，ARIB,MPEG,EBUなどの標準化活動に従事。現在，放送技術研究所人間・情報科学研究部主任研究員。博士（情報学）。

住吉英樹（すみよしひでき）

1980年入局。広島放送局を経て，1984年から放送技術研究所にて，コンピューターを応用した番組制作システム，メタデータ制作システムの研究に従事。現在，放送技術研究所人間・情報科学研究部専任研究員。博士（工学）。

枠組みMPF（メタデータ制作フレームワーク）について，その仕様を紹介した。MPFの利点は，放送局や映像コンテンツ提供事業者にとっては，新規技術をモジュールとして組み込むことによって，常に最新のメタデータ制作環境が活用できるようになることである。処理技術を提供する事業者においては，個別に大きなシステムを構築する必要がなく，ある処理に特化した独自技術をモジュールとして供給することができ，効率のよい環境であると考えている。また，学術的には，個々の研究成果がモジュールになり，研究室などでの技術の蓄積・継承・利用が容易になるほか，研究の効率を更にあげることができると考えている。

MPFはメタデータ制作という目的を通して，さまざまなメディア解析技術，情報処理技術を発展させることに貢献するものであり，今後，更に実践的な目標を定めて，賛同していただける研究機関と連携して研究開発を進めていく予定である。

参考文献

1）メタデータ制作フレームワーク，http://www.nhk.or.jp/strl/mpf/
2）ISO/IEC 15938, “Information technology − Multimedia content description interface”
3）XQuery 1.0, An xml query language, http://www.w3.org/TR/xquery/
4）eXist：Open source native XML database, http://exist.sourceforge.net/
5）Information Grand Voyage project, http://www.igvpj.jp/
6）P/SCAIE project, http://tech.ebu.ch/groups/pscaie/

藤井真人（ふじいまひと）

1983年入局。札幌放送局を経て，現在，放送技術研究所人間・情報科学研究部主任研究員。この間，CMUに半年間滞在，およびATR人間情報通信研究所に出向。神経回路モデル，視覚情報処理，画像認識，映像検索などの研究開発に従事。博士（情報科学）。

柴田正啓（しばたまさひろ）

1981年入局。新潟放送局，放送技術研究所，技術局，放送技術局を経て，現在，放送技術研究所人間・情報科学研究部部長。情報検索，画像データベース，映像ハンドリング技術，番組制作システムなどの研究開発に従事。博士（情報学）。

八木伸行（やぎのぶゆき）

1980年入局。甲府放送局，放送技術研究所，技術局，編成局を経て，現在，放送技術研究所研究企画部部長。画像・映像・メディア情報処理，コンピューターアーキテクチャー，コンテンツ制作技術，デジタル放送などの研究開発に従事。博士（工学）。

報告

逐次的な判定手続きに基づくショット境界の高速検出手法

河合吉彦　住吉英樹　八木伸行

Fast Method of Shot Boundary Detection Based on Sequential Decision Procedure

Yoshihiko KAWAI, Hideki SUMIYOSHI and Nobuyuki YAGI

要約

映像からカメラの切り替え点を検出するショット境界検出は連続的なデータである映像を小さな構成単位（ショット）に分割する技術であり，映像解析における最も基本的な処理の１つである。本稿では，従来手法と同程度の検出精度を維持したまま，より高速にショット境界を検出する手法として，ショット境界の可能性が低いフレームに対しては処理を省略し，可能性の高い部分についてだけさまざまな特徴量を逐次的に算出する手法を提案する。実際の放送映像に対する実験では，再現率が90.4%，適合率が92.8%という良好な結果が得られた。また，約425分のテストデータに対する処理時間は208秒（MPEG－１のデコード時間を除く）であり，実時間の約１/123という高速な処理を実現した。

ABSTRACT

Shot boundary detection is defined as processing of dividing video data into short video segments based on shot boundaries which are points in time when TV cameras are switched. Shot boundary detection is one of the most fundamental processes in video analysis. In this paper, we propose a novel method which enables precise, high-speed detection by omitting the processing of frames that are clearly not shot boundaries, and by analyzing various features sequentially only for the parts of the video that are likely to contain shot boundaries. An evaluation on actual broadcast video resulted in a recall rate of 90.4% and a precision rate of 92.8% on average. About 425 minutes of test data was processed in 208 seconds (excluding the MPEG-1 decoding time), or 1/123rd of the real time.

1．まえがき

放送映像は連続的なデータなので，映像解析を行うためには，まず，映像を小さな構成単位に分割する必要がある。構成単位としては，テキストにおける単語のように意味的にまとまりを持つものが望ましいが，映像解析では，一般的に，映像の物理的な特徴に基づいて分割が可能であるショット（１台のカメラで連続的に撮影されたフレーム列）が利用される。放送映像をショットに分割するためには，ショットとショットのつなぎ目であるショット境界を検出する必要がある。この検出処理は映像解析における最も基本的な処理の１つであり，ショット境界検出と呼ばれている。

ショット境界は前のショットから次のショットへ瞬時にショットを切り替える瞬時切り替えと，複数フレームにわたって徐々に切り替える漸次切り替えに大別できる。瞬時切り替えはカットとも呼ばれ，映像編集において最も頻繁に使用される切り替え方法である。漸次切り替えは２つのフレームの合成の割合を徐々に変化させるディゾルブやフェードなどに分類できる。ディゾルブは前のショットの画面全体が徐々に消えていき，次のショットの画面全体が徐々に現れてくるような切り替え方法である。また，フェードはディゾルブの１種であり，前のショットあるいは次のショットの全体が黒であるものを指す。

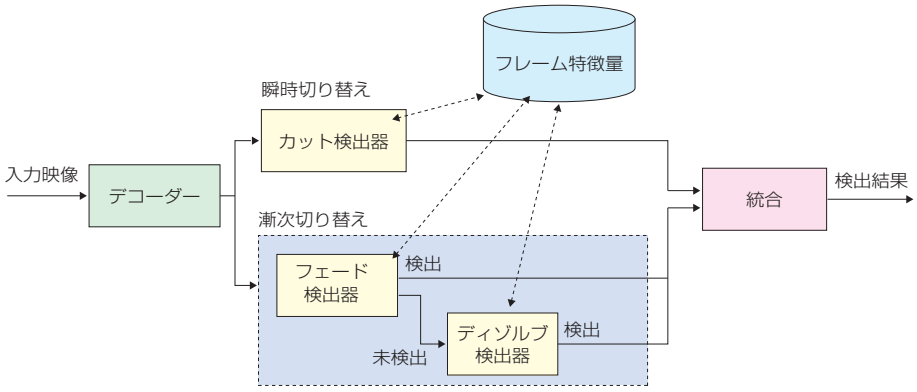
ショット境界は近接フレーム間の類似度に基づいて検出することができる。これは，２つのフレームが同一のショットに属している場合にはフレーム間の差分は小さく，２つのフレームの間にショット境界が存在する場合にはフレーム間の差分が大きいという特性に基づいている。フレーム間の差分を表す特徴量としては，色や輝度のヒストグラム差分[1]，エッジの変化量[2]，動きベクトルの変化量[3]，階調値の分散[4]，画素間の相互情報量[5]など，さまざまな特徴量が提案されている。しかし，１種類の特徴量だけでは，カメラや被写体の動きによる変化と，ショット切り替えによる絵柄の変化とを精度よく区別することは難しい。そこで，近年，複数の特徴量を組み合わせる手法が提案されている[6][7][8][9]。エッジや色などの特徴を総合的に判断することで，単一の特徴量に基づく手法と比較して，高い検出精度を得ることができる。しかし，これらの手法では，入力映像の各フレームに対してさまざまな特徴量を計算する必要があり，計算コストが高くなるという問題がある。

そこで，本稿では，従来手法と同程度の検出精度を維持したまま，計算コストを削減する手法を提案する。提案手法では，従来手法のように，すべてのフレームに対して多くの特徴量を算出するのではなく，フレームごとに画像特徴量の算出とその特徴量に基づくショット境界の判定処理を逐次実行する。ショット境界ではないと判定した時点でそのフレームに対する処理を終了し，次のフレームの処理に移る。放送映像の大部分はショット境界以外のフレームによって占められており，これらのフレームに要する処理時間を短縮することで，全体の計算コストを大幅に削減することができる。

２章で提案手法の概要を説明し，３章，４章および５章で検出器の詳細について説明する。６章では提案手法を検出精度と計算コストの両面から評価する。７章で全体のまとめと課題を述べる。

2．逐次判定によるショット境界検出

１図に提案手法の概要を示す。まず，入力された放送映像をデコードし，フレーム画像を取得する。次に，カット検出器，フェード検出器およびディゾルブ検出器を利用してショット境界を検出する。ショット境界の検出精度を上げるために，瞬時切り替え（カット）と漸次切り替え（フェード，ディゾルブ）の検出器を並列に配置する。なお，漸次切り替えはフェード，ディゾルブの順に処理する。フェードが検出された場合には，ディゾルブの検出処



1図　ショット境界検出処理の概要

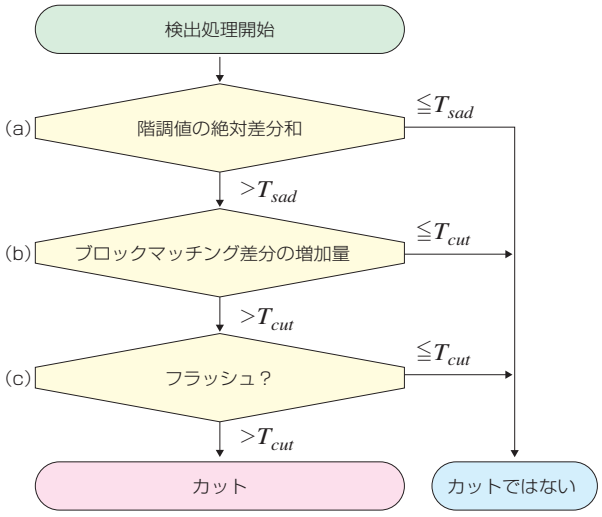
理を実行しないことで，誤検出の軽減と処理の高速化を図る。また，各検出器で算出されるさまざまなフレーム特徴量を３種類の検出器で共有し，同じ特徴量を重複して計算しないようにする。最後に，複数の検出器で検出されたショット切り替えを統合し，最終的な検出結果を出力する。統合処理では，ディゾルブと検出された区間内にカットが検出された場合や１つのディゾルブ区間と考えられる区間に複数のディゾルブ区間が検出された場合などを最終的にどのような検出結果とするかを経験則に基づいて判断する。以下，本稿の主題である３種類の検出器の詳細について述べる。

３．カット検出器

２図にカット検出器の検出手順を示す。カットは切り替えの前後でフレーム画像が大きく変化する。そのため，まず始めに，隣接フレームにおける赤（R），緑（G），青（B）の各階調値の絶対差分和d_{sad}を算出して，カットであるかどうかを判定する（２図（a））。d_{sad}は（１）式で与えられる。

$$d_{sad}(f_{i-1}, f_i) = \frac{1}{|\boldsymbol{F}|}\sum_{\boldsymbol{r}\in\boldsymbol{F}}|f_{i-1}(\boldsymbol{r}) - f_i(\boldsymbol{r})| \quad \text{（１）}$$

ここで，$f_i(\boldsymbol{r})$はi番目のフレームの座標$\boldsymbol{r}$における画素値を表す。また，$\boldsymbol{F}$はフレーム全体の画素を，$|\boldsymbol{F}|$は画素の総数を表す。実際にはR，G，Bごとに階調値の絶対差分和を求めるが，式が煩雑になるので，ここでは簡略化した式を示した。d_{sad}はフレームの変化に対して敏感に反応するので，誤検出が多発する可能性があるが，未検出は少ないという特徴がある。d_{sad}がしきい値T_{sad}以下である場合には，カットではないと判定して，そのフレームに対する処理を終了し，次のフレームの処理に移る。



２図　カット検出の手順

次に，より精度の高いブロックマッチング差分d_{bm}を算出する（２図（b））。ブロックマッチングでは，フレームの映像を空間的に複数の矩形（くけい）状のブロックに分割し，各ブロックの映像と前のフレームの各ブロックの映像との差（以下，ブロック間コストと呼ぶ）が最小となる位置を探索する。ブロック間コストの算出には，カメラの動きに頑健な輝度ヒストグラムの絶対差分和を用いる。ブロック間の最小コストがしきい値を超えるブロックの数を総数で正規化した値をd_{bm}とする。なお，輝度ヒストグラムはR，G，Bのそれぞれに対して算出する。ショット境界の判定には，カメラの激しい動きなどによる誤検出を防ぐために，ブロックマッチング差分の増加量を用いる。カットと判定する条件を（２）式に示す。

$$d_{bm}(f_{i-1}, f_i) - d_{fbm}(f_{i-2}, f_{i-1}) > T_{cut} \quad \text{（２）}$$

ここで，d_{fbm}は前のフレームに対して算出するフレーム間差分であり，（３）式で与えられる。

$$d_{fbm}(f_{i-2}, f_i) = \begin{cases} 0 & if \quad d_{sad}(f_{i-2}, f_{i-1}) \leqq T_{sad} \\ d_{bm}(f_{i-2}, f_{i-1}) & else \end{cases} \quad \text{（３）}$$

本稿では，d_{fbm}を高速ブロックマッチング差分と呼ぶ。（２）式を用いてカットでないと判定された場合には，そのフレームに対する処理を終了して，次のフレームの処理に移る。

最後に，カメラのフラッシュによって引き起こされる誤検出を判定する（２図（c））。フラッシュが発光されると，最初の数フレームの間だけ輝度が上昇し，その後，元に戻る。そこで，フレームf_iからフレームf_{i+N_f}の同じ位置における画素の最小輝度値で構成される画像f_i^*を合成する。（４）式にf_i^*の生成式を示す。

$$f_i^*(\boldsymbol{r}) = \min\left(f_i(\boldsymbol{r}), f_{i+1}(\boldsymbol{r}), \cdots, f_{i+N_f}(\boldsymbol{r})\right) \quad \text{（４）}$$

ここで，N_fはフラッシュの発光時間よりも長く設定する。f_iがフラッシュによって輝度上昇した場合には，f_{i-1}とf_i^*は類似すると考えられる。そこで，f_{i-1}とf_i^*の高速ブロックマッチング差分d_{fbm}を算出し，$d_{fbm}(f_{i-2}, f_{i-1})$との差分がしきい値$T_{cut}$以下であればフラッシュによる誤検出，$T_{cut}$より大きければカットと判定する。カットと判定する条件は（５）式である。

$$d_{fbm}(f_{i-1}, f_i^*) - d_{fbm}(f_{i-2}, f_{i-1}) > T_{cut} \quad \text{（5）}$$

カット検出器では，前述のすべての判定処理を通過したフレームをカットと判定する。

4．フェード検出器

漸次切り替えでは，まず，ディゾルブよりも検出精度の高いフェード検出を行う。先にも述べたように，フェードはディゾルブの１種で，切り替え区間の開始フレームあるいは終了フレームの画面全体が黒（以下，黒フレームと呼ぶ）である。フェード検出器の検出手順を３図に示す。

最初に，入力フレームが黒フレームであるかどうかを判定する。黒フレームの判定には，フレーム全体の平均輝度値と，輝度値の低い画素の割合を利用する。平均輝度値がしきい値T_{blum}以下であり（３図（a)），暗い画素の割合がしきい値T_{barea}より大きい場合に（３図（b））黒フレームと判定する。黒フレームでないと判定した場合には，そのフレームに対する処理を終了し，ディゾルブの検出に移る。

黒フレームと判定された場合には，現フレームを終点とするフェードアウトであるか，または，現フレームを始点とするフェードインであるかの判定をそれぞれ行う。フェードインあるいはフェードアウトによる切り替え区間は，複数フレームにわたって連続して輝度が単調増加あるいは単調減少するかによって検出する（３図（c））。単調増加あるいは単調減少している画素の割合がしきい値T_{fade}より大きい場合にはフェードの可能性があると判定する。

輝度の低い物体の面積が徐々に増加あるいは減少する場合にも，単調増加あるいは単調減少する画素の割合がしきい値T_{fade}よりも大きくなることがある。そこで，輝度の低い物体の面積が徐々に変化する場合とフェードとを区別するために，隣接フレーム間の類似度（余弦類似度）を算出し，その値がしきい値T_{fsim}以下である場合をフェードの可能性があると判定する（３図（d））。３図（c）および３図（d）が連続的に成り立つ区間をフェード区間として検出する。なお，余弦類似度がしきい値T_{fsim}より大きい場合には，フレーム番号を＋１（または，－１）して，輝度が単調増加あるいは単調減少する画素の割合がしきい値T_{fade}以下になるまでループする。

最後に，検出されたフェード区間の長さが十分に長い場合をフェードと判定する（３図（e））。なお，３図でフェード開始フレームと検出された場合においても，その次のフレームを３図の最初の入力フレームとして処理を行い，フェードの検出精度を上げる。すなわち，理想的なフェード検出器であれば，１つのフェード区間に対して，複数回，フェード区間であると判定することになる。

5．ディゾルブ検出器

フェード検出器でフェード未検出のフレームに対して，ディゾルブ検出を行う。本稿では，（６）式でディゾルブのモデルを表現する。

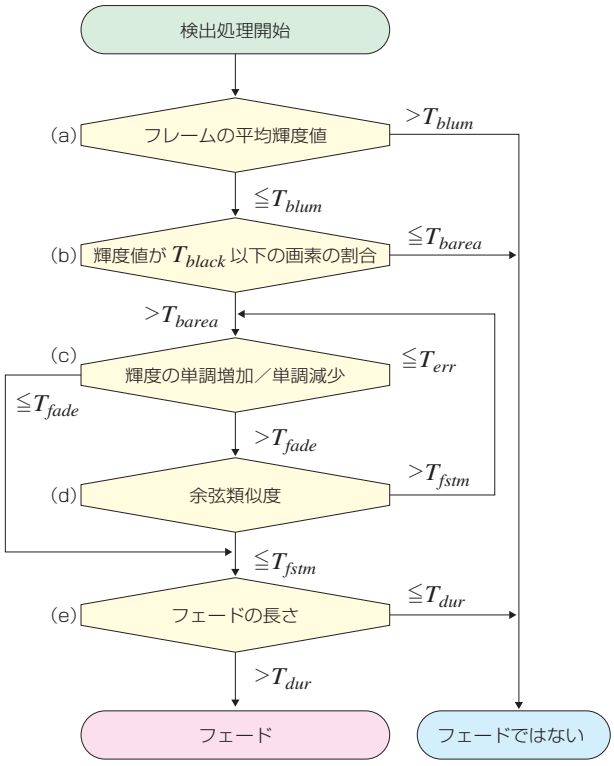
$$f_i(\boldsymbol{r}) = (1-\beta)\cdot f_{db}(\boldsymbol{r}) + \beta \cdot f_{de}(\boldsymbol{r}),\quad 0 \leqq \beta \leqq 1 \quad \text{（6）}$$

ここで，f_{db}はディゾルブによる切り替えの開始フレームを，f_{de}は終了フレームを表す。また，βは開始フレームでは０であり，終了フレームでは１である。

４図にディゾルブ検出器の検出手順を示す。ディゾルブでは複数のフレームにわたって映像が徐々に変化するので，カットと同じ手法で検出することは難しい。

まず，現フレームf_iとN_dフレーム前のフレームf_{i-N_d}との絶対差分和$d_{sad}(f_{i-N_d}, f_i)$を用いて，ディゾルブの可能性がある区間を検出する（４図（a））。$d_{sad}(f_{i-N_d}, f_i)$がしきい値T_{sad}以下の場合には，ディゾルブではないと判定し，そのフレームに対する処理を終了し，次の入力フレームを待つ。

次に，（６）式のモデルに基づく２種類の特徴量を用いてディゾルブの候補区間を検出する。第１の特徴量は単



3図　フェード検出の手順

調に変化する画素の割合である。フレームが（6）式に従って単調に変化するとき，フレームの各画素の値は単調に増加あるいは単調に減少する。そこで，第1の特徴量として（7）式を定義する。

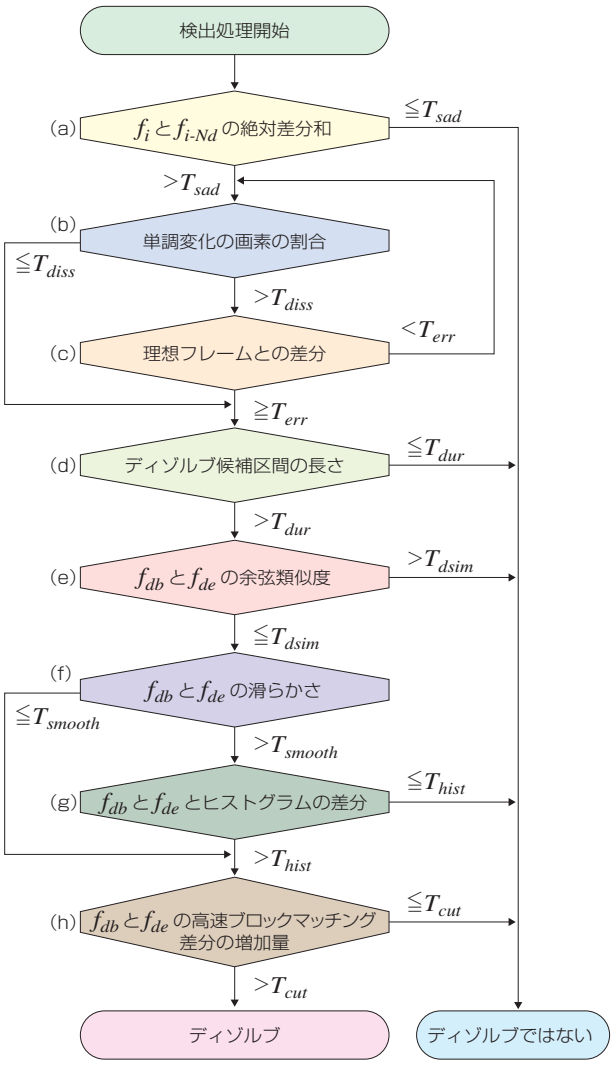
$$g_{diss}(f_{i-1}, f_i, f_{i+1}) = \frac{1}{|F|}\sum_{r\in F}\delta_{mono}(f_{i-1}(r), f_i(r), f_{i+1}(r)) \quad \text{——— (7)}$$

ここで，

$$\delta_{mono}(f_{i-1}(r), f_i(r), f_{i+1}(r)) = \begin{cases} 1 & if \quad (f_{i-1}(r) > f_i(r) > f_{i+1}(r)) \\ & or \quad (f_{i-1}(r) < f_i(r) < f_{i+1}(r)) \\ 0 & else \end{cases} \quad \text{——— (8)}$$

である。

第2の特徴量は理想的なディゾルブとの差分である。



4図　ディゾルブ検出の手順

（6）式においてβが一定の速度で変化する場合には，（9）式が成り立つ。

$$f_i(r) = \frac{f_{i-1}(r) + f_{i+1}(r)}{2} \quad \text{——— (9)}$$

そこで，(10)式で算出される絶対差分和を第2の特徴量とする。

$$g_{err}(f_{i-1}, f_i, f_{i+1}) = \frac{1}{|F|}\sum_{r\in F}\left|\frac{f_{i-1}(r) + f_{i+1}(r)}{2} - f_i(r)\right| \quad \text{——— (10)}$$

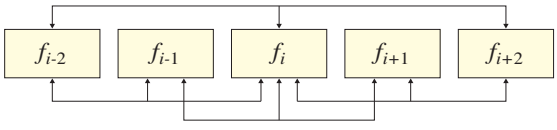
2つの特徴量を用いた候補区間の具体的な判定方法を説明する。(7)式および(10)式のような現フレームとその前後のフレームに基づく特徴量ではカメラや被写体の動きによる誤検出が多発する。そこで，提案手法では，5図に示す4種類のフレームの組み合わせに対して特徴量を算出する。(11)式と(12)式が同時に成り立つ場合には，フレーム番号を＋1してループし，(11)式と(12)式が同時に連続して成り立つ区間（以下，ディゾルブ候補区間と呼ぶ）の長さを求める（4図（b），（c））。なお，(11)式と(12)式が1度も同時に成り立たない場合には，ディゾルブ候補区間の長さは0である。

$$g_{diss}(f_{i-2}, f_{i-1}, f_i) + g_{diss}(f_{i-1}, f_i, f_{i+1}) + g_{diss}(f_i, f_{i+1}, f_{i+2}) + g_{diss}(f_{i-2}, f_i, f_{i+2}) > 4T_{diss} \quad \text{——— (11)}$$

$$g_{err}(f_{i-2}, f_{i-1}, f_i) + g_{err}(f_{i-1}, f_i, f_{i+1}) + g_{err}(f_i, f_{i+1}, f_{i+2}) + g_{err}(f_{i-2}, f_i, f_{i+2}) < 4T_{err} \quad \text{——— (12)}$$

次に，検出されたディゾルブ候補区間の長さが十分であるかどうかを判定する。ディゾルブ候補区間の長さがしきい値T_{dur}以下の場合には，ディゾルブではないと判定し，そのフレームに対する処理を終了する（4図（d））。

アイリスの変動などによって輝度値が時間的に滑らかに変化する場合にも(11)式と(12)式が成り立ち，誤検出となる場合がある。そこで，候補区間の開始フレームf_{db}



5図　輝度の単調変化に基づくディゾルブ検出

と終了フレーム f_{de} の類似度を算出し，それがしきい値 T_{dsim} よりも大きい場合には誤検出と判定し，そのフレームに対する処理を終了する（4図（e））。

また，輝度値が空間的に滑らかに変化するような画像が一定の方向に移動した場合にも誤検出となる場合がある。そこで，f_{db} および f_{de} が滑らかな背景を持つ画像かどうかを判定する（4図（f））。判定には，(13)式の特徴量を利用する。

$$g_{smooth}(f) = \frac{1}{4|\boldsymbol{F}|}\sum_{(x,y)\in\boldsymbol{F}}\{s_h(f)+s_v(f)+s_{dr}(f)+s_{dl}(f)\} \quad \text{(13)}$$

ここで，

$$s_h(f) = \delta_{mono}\,(f(x-1,y), f(x,y), f(x+1,y)) \quad \text{(14)}$$

$$s_v(f) = \delta_{mono}\,(f(x,y-1), f(x,y), f(x,y+1)) \quad \text{(15)}$$

$$s_{dr}(f) = \delta_{mono}\,(f(x-1,y-1), f(x,y), f(x+1,y+1)) \quad \text{(16)}$$

$$s_{dl}(f) = \delta_{mono}\,(f(x-1,y+1), f(x,y), f(x+1,y-1)) \quad \text{(17)}$$

である。

$g_{smooth}(f_{db})$ または $g_{smooth}(f_{de})$ がしきい値 T_{smooth} より大きい場合には，更に，f_{db} と f_{de} の輝度ヒストグラム差分 d_{hist} を算出し，この値がしきい値 T_{hist} 以下である場合には誤検出と判定し，そのフレームに対する処理を終了する（4図（g））。

最後に，f_{db} と f_{de} の高速ブロックマッチング差分 d_{fbm} を計算し，d_{fbm} の増加量がしきい値 T_{cut} よりも大きい場合にディゾルブと判定する（4図（h））。

ディゾルブの中には，ごくまれに数秒間にわたる長い切り替え時間を持つものがある。このような長いディゾルブでは，隣接フレーム間において画像がほとんど変化しないので，前述の処理では検出が難しい。そこで，通常のディゾルブ検出を実施した後，(11)式と（12)式におけるフレーム間隔を N_{ld} 倍に設定し，再度，検出処理を実行し，切り替え区間の長いディゾルブを検出する。

6．評価実験

実験にはTRECVID（Text REtrieval Conference − VIDeo retrieval evaluation）2007の映像データセットを利用した。TRECVIDは米国国立標準技術研究所（NIST：National Institute of Standards and Technology）が主催している映像検索やコンテンツ解析に関する国際的なベンチマークワークショップである[10]。映像データはオランダで放送されたテレビ番組であり，内容は主にドキュメンタリーや子供番組，教育番組などである。映像データの合計長は425分である。また，フレームサイズは352×288ピクセル，フレームレートは25 fpsであり，映像はすべてMPEG－1形式でエンコードされている。

本手法で利用するしきい値はさまざまなジャンルの放送映像（ドラマ，スポーツ，ドキュメンタリーなど）を用いた予備実験を行って決定した。具体的には，しきい値をさまざまに変化させて，検出精度が良好な値を実験に用いた。検出精度の評価には，再現率（*recall*），適合率（*precision*），F値（$F-measure$）を用いた。再現率および適合率は（18)式で与えられる。

$$recall = \frac{N_{both}}{N_g}, \quad precision = \frac{N_{both}}{N_o} \quad \text{(18)}$$

ここで，N_g は映像に含まれるショット境界の総数，N_o は提案手法で検出したショット境界の総数，N_{both} は映像に含まれるショット境界のうち提案手法で検出できた数を表す。再現率は未検出の少なさを，適合率は誤検出の少なさを表す指標である。また，(19)式で示すF値は，トレードオフの関係にある再現率と適合率を統合した指標で，未検出および誤検出の両方を考慮した全体的な精度を表す。

$$F-measure = \frac{2\cdot recall\cdot precision}{recall+precision} \quad \text{(19)}$$

6.1　検出精度の評価

実験結果を**1表**に示す。ショット境界検出の再現率は90.4%，適合率は92.8%であり，非常に良好な結果であった。そのうち，瞬時切り替えだけの検出精度は，再現率が93.0%，適合率が96.2%であった。また，漸次切り替えだけでは再現率，適合率共に60%程度であった。

未検出および誤検出となった個所を調査した。未検出については，輝度が不安定に変動する番組（フィルム撮影による映像劣化の激しい番組）において，瞬時切り替えの再現率が約50%と低かった。定常的な輝度変動の影響によってフレーム間差分が常に大きくなり，(3)式で示されるフレーム間差分の増加量がしきい値以上とならず，未検出が多発したと考えられる。そのほか，類似したショット間の切り替えを見落とす場合があった。これに対処するためには，輝度ヒストグラムのビン数*1やしきい値を動的に変化させるなどの検討が必要である。漸次切り替えにつ

1表　ショット境界の検出精度

	再現率	適合率	F値
瞬時切り替え	93.0%	96.2%	0.946
漸次切り替え	61.2%	63.0%	0.604
合　計	90.4%	92.8%	0.915

2表　ショット境界検出の処理時間（実験データ：約425分）

処理内容	処理時間	実時間に対する比率
デコード処理	1,478.5秒	1/17
セグメンテーション処理	208.0秒	1/123
合　計	1,686.5秒	1/15

いては，切り替え区間内でカメラや被写体が動く場合に，ディゾルブが検出できない事例があった。これは，提案手法のディゾルブ検出器では，切り替え区間内のフレームは静止画であることを仮定しているためである。動きや変化がある場合への対応が必要である。

誤検出については，サイズの大きなスーパーインポーズの出現や，カメラレンズの前を物体が横切るシーンにおいて，カットと判定する場合があった。また，被写体の動きなどによって，輝度の時間変化がディゾルブと類似した特徴となり，誤検出となる場合があった。これについては，エッジ特徴や周波数特徴などを判定条件に加えるなど，ディゾルブと被写体の動きとを区別する処理が必要である。

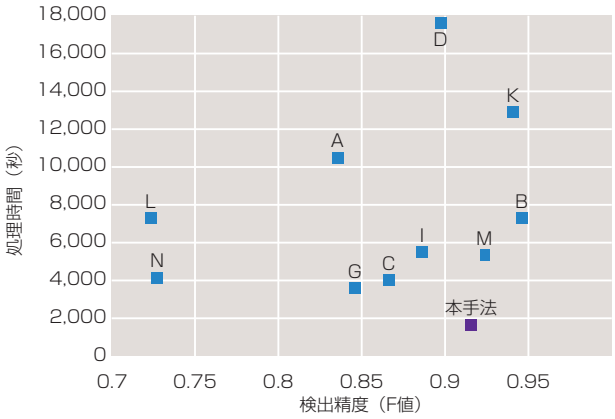
6.2　処理時間の検証

実験には，CPUがIntel Core 2 Duo E6600 2.4 GHz，メモリーが2 Gバイトの計算機を使用した。425分の実験データに対する処理時間を**2表**に示す。2表における「デコード処理」はMPEG－1映像を復号してフレーム画像を取得するのに要した時間であり，「セグメンテーション処理」はフレーム画像を解析してショット境界を判定するのに要した時間である。

提案手法における処理時間の合計は1,686.5秒（約28分）であり，実時間の約1/15という高速な処理であった。また，デコード時間を除いた処理時間は208秒であり，実時間の約1/123であった。処理時間のうちの大部分は映像のデコード処理に要しており，これを最適化することで，更なる高速化が期待できる。

6.3.　既存手法との比較

既存手法との比較結果を**6図**に示す。比較に用いた手法はTRECVID 2007のショット境界検出課題に提出された手法[11]である。グラフの横軸がF値，縦軸が処理時間であり，右下にある手法ほど高速で高精度な手法であることを表している。提案手法の検出精度はK，B，Mの手法に次いで4番目の精度であった。提案手法のF値は0.915であり，他の手法と比較してもそん色のない結果であった。また，処理時間については，CPUの性能に多少の差はあったが，提案手法が最も高速で，処理時間は他の手法の



6図　検出精度と処理時間の比較

1/2から1/360程度であった。検出精度が上位の3手法については，提案手法よりもF値は0.01から0.03程度高かったが，処理時間は提案手法の3～8倍程度であった。提案手法は処理時間と検出精度の両方が要求されるアプリケーションにおいて有効な手法と言える。

7．あとがき

高精度で高速にショット境界が検出可能な手法を提案した。提案手法では，ショット境界の可能性が低いフレームについては処理を省略し，可能性が高い部分についてだけさまざまな特徴量を逐次的に算出することで，検出精度と計算コストを両立させた。実際の放送映像を利用した実験では，再現率が90.4%，適合率が92.8%という良好な結果が得られた。また，約425分の放送映像に対する処理時間は208秒（MPEG－1のデコード時間を除く）であり，実時間の約1/123という高速な処理が実現できた。

今後の検討課題としては，各処理におけるしきい値の決定方法がある。学習データから最適な組み合わせを決定するような手法についての検討が必要である。

本稿は電子情報通信学会論文誌に掲載された以下の論文を元に加筆・修正したものである。
河合，住吉，八木：“逐次的な特徴算出によるディゾルブ，フェードを含むショット境界の高速検出手法，”電子情報通信学会論文誌，Vol.J91−D, No.10, pp.2529−2539（2008），

＊1　数値がとり得る範囲を幾つに分割するかという値。

参考文献

1） H. Zhang, A. Kankanhalli and S.W. Smoliar：“Automatic Partitioning of Full−Motion Video,” Multimedia Systems, Vol.1, No.1, pp.10−28（1993）

2） R. Zabih, J. Miller and K. Mai：“A Feature−Based Algorithm for Detecting and Classifying Production Effects,” Multimedia Systems, Vol.7, No.2, pp.119−128（1999）

3） 鈴木，中嶋，坂野，三部，大塚：“動き方向ヒストグラム特徴を用いた映像データからのカット点検出法,” 信学論（D），Vol.J86−D2, No.4, pp.468−478（2003）

4） J. Nam and A.H. Tewfik：“Detection of Gradual Transitions in Video Sequences Using B−Spline Interpolation,” IEEE Trans. on Multimedia, Vol.7, No.4, pp.667−679（2005）

5） Z. Cernekova, I. Pitas and C. Nikou：“Information Theory−Based Shot Cut/Fade Detection and Video Summarization,” IEEE Trans. on Circuits and Syst. Video Technol, Vol.16, No.1, pp.82−91（2006）

6） X. Gao and X. Tang：“Unsupervised Video−Shot Segmentation and Model−Free Anchorperson Detection for News Video Story Parsing,” IEEE Trans. Circuits and Syst. Video Technol, Vol.12, No.9, pp.765−776（2002）

7） C.−W. Ngo：“A Robust Dissolve Detector by Support Vector Machine,” Proc. ACM Multimedia, pp.283−286（2003）

8） K. Matsumoto, M. Naito, K. Hoashi and F. Sugaya：“SVM−Based Shot Boundary Detection with a Novel Feature,” Proc. IEEE ICME’06, pp.1837−1840（2006）

9） Z. Liu, E. Zavesky, D. Gibbon, B. Shahraray and P. Haffner：“AT＆T Research at TRECVID 2007,” Proc. TRECVID Workshop, pp.19−26（2007）

10） A.F. Smeaton, P. Over and W. Kraaij：“Evaluation Campaigns and TRECVid,” Proc. ACM MIR’06, pp.321−330（2006）

11） National Institute of Standards and Technology：“TRECVID 2007 Runs and Detailed Results：Shot Boundary Determination,” Proc. TRECVID Workshop, pp.A−1−A−7（2007）

河合吉彦（かわいよしひこ）

2001年入局。放送技術局を経て，2005年から放送技術研究所にてメディア処理の研究に従事。現在，放送技術研究所人間・情報科学研究部に所属。博士（工学）。

住吉英樹（すみよしひでき）

1980年入局。広島放送局を経て，1984年から放送技術研究所にてコンピューターを応用した番組制作システム，メタデータ制作システムの研究に従事。現在，放送技術研究所人間・情報科学研究部専任研究員。博士（工学）。

八木伸行（やぎのぶゆき）

1980年入局。甲府放送局，放送技術研究所，技術局，編成局を経て，現在，放送技術研究所研究企画部部長。画像・映像・メディア情報処理，コンピューターアーキテクチャー，コンテンツ制作技術，デジタル放送などの研究開発に従事。博士（工学）。

報告

投球の次ショットに重きを置いたシーンのシンボル列化による野球放送映像プレー種分類

望月貴裕　藤井真人　八木伸行　篠田浩一†

Play–Classification of Baseball Broadcast Video Scenes Based on Symbol–Sequenced Scene Focusing on Post –Pitch Shot

Takahiro MOCHIZUKI, Mahito FUJII, Nobuyuki YAGI and Kouichi SHINODA

要約

野球放送の映像シーンを数種類のプレー種（本塁打，シングルヒット，四球など）へ自動的に分類する手法について述べる。提案手法は，画像の特徴や動きの情報を持つ矩形（くけい）集合を用いて映像区間を簡略化して表現する技術に基づいている。簡略化処理の基本単位はショットである。更に，各シーンの投球ショットの次ショットは，プレー種を区別するための重要な情報を含む映像区間なので，ショットを固定長で分割した部分ショットを簡略化の処理単位とする。学習用の野球映像シーンを，各ショットおよび部分ショットを簡略化したシンボル列で表現し，プレー種ごとの離散隠れマルコフモデル（離散HMM）でそれらのシンボル列を学習する。学習済みのHMMを用いた尤度（ゆうど）計算を行い，プレー種が未知のシーンにプレー種を割り当てる。本稿では，更に，MLB（Major League Baseball）放送の映像を用いた実験を行い，高精度でシーンを分類できることを示す。

ABSTRACT

This paper describes a method for automatically classifying baseball video scenes into play–classes（i.e., homerun, single, walk, etc.）. Our method is based on a technique to simplify a video interval using a set of rectangles with image features and motion information. The basic unit for simplification is a shot. For the second shot of each scene that includes significant information for play–classification, a partial shot generated by dividing the shot is used as a processing unit. The scenes used for training are expressed as sequences of symbols based on the simplified data for shots and partial shots. “Play–class–unknown” baseball scenes are assigned one of the play–classes by using discrete hidden Markov models that have been trained with the training symbol sequences for each kind of play–class. An experiment using videos of seven Major League Baseball games produced good results, demonstrating that this method can automatically classify scenes with high accuracy.

†東京工業大学　Tokyo Institute of Technology

1．まえがき

ソフトウエア，ハードウエアの高度化により画像や映像を電子化して扱うことが容易な時代となった。また，符号化技術の発展や記録媒体の大容量化により，家庭のPC（パーソナルコンピューター）やハードディスクレコーダーに大量の映像データが蓄積可能となった。過去のテレビ放送などの貴重な映像資産を大規模に保管して，その一部を公開するアーカイブス施設[1]の運用も進められている。

更に，インターネットを中心とした通信環境が整備され，通信網上の広い範囲に存在するさまざまな映像データにユーザーが容易にアクセス可能となっている。放送の世界においても，インターネットを使ったオンデマンド配信サービス[2]や，素材の提供サービス[3]などが開始されており，今後，ユーザーの増加が期待されている。

そのような流れの中で，大量の映像を内容に基づいて管理・検索する技術が求められている[4]。特に，スポーツ映像では生起するプレー種が比較的区別しやすく，ユーザーの「何が見たいか」という要求も明確なので，検索技術の適用が強く望まれており，スポーツ映像のシーンを対象とした内容による自動分類，ハイライト抽出，映像要約などのさまざまな研究が行われている。

馬場口らはアメリカンフットボールを処理対象として，映像に付与されているクローズドキャプション＊1と簡易な色特徴を用い，各シーンを数種類のプレー種へ分類する技術を提案している[5]。馬場口らの手法は，クローズドキャプションの正確さに強く依存しており，その信頼度が低い映像あるいはまったく付与されていない映像への適用は難しい。渡辺らは野球映像を対象として，フレーム画像を空間的に分割したサブブロックごとの動きベクトルのマハラノビス距離＊2の学習に基づく方法で，投球ショットとそれ以外のショットに分類している[6]。しかし，映像から抽出する画像情報量および学習データ量が十分ではなく，分類対象として扱うことのできるプレー種が大きく制限されている。Ekinらはフレーム画像中の芝生領域の割合およびペナルティーボックスの位置を特徴量としてサッカー映像のシーンの分類を試みている[7]。しかし，簡易な特徴のマッチング処理に基づく分類なので分類精度が低く，それを向上させるためには，学習処理を強化する必要がある。

隠れマルコフモデル（HMM）＊3は時系列で推移する事象の解析に非常に適したモデルで，映像から抽出した特徴の推移に基づくHMMを利用したさまざまなシーン分類，イベント認識手法が提案されている。Xuらはブロック単位の動きデータを画像特徴としてHMMに学習させ，バスケットボールの映像のイベント認識を行っている[8]。しかし，Xuらの手法は，大部分が水平方向のカメラの動きで構成されるバスケットボール特有の条件を満たす映像を処理対象としている。Xieらはフレーム画像の主要色領域の比率や動き情報の推移を学習したHMMを用いてサッカー映像のシーン分類を試みているが[9]，「プレーシーン」と「ブレイクシーン」の２種類のみの判別にとどまっている。Reaらはテニス映像を対象として，コート上のラインおよび選手の位置をフレーム画像から抽出し，その推移を学習したHMMを用いて，得点シーンなどのシーン分類を行っている[10]。しかし，プレー中の映像がテニスコートのルーズショットだけで構成されていることを想定しており，選手のアップや観客席の映像などが挿入される一般の放送映像への適用は困難である。

HMMを用いたスポーツ映像のシーン分類では，野球放送の映像を対象とした手法が多く提案されている。野球放送の映像はホームラン，ツーベース，シングルヒット，四球などのプレー種ごとにカメラの切り替えなどによる画の流れのパターンが明確に異なる傾向があり，HMMを用いた学習および認識処理に非常に適しているからである。

各シーンにおけるフレーム単位での画像の特徴の推移を学習したHMMを用いてシーンのプレー種を判定する手法が提案されている[11]。フレームを推移の単位とするこの手法は，時間方向の伸縮，例えば，投手がいったんプレートを外して投球間隔が伸びる場合などに対しては影響を受けやすい。また，カメラのスイッチングミスや物体の横切りなど，想定外の映像が短時間挿入されると分類精度が低下する。また，HMMのモデルとして連続HMM＊4を採用しているので，学習時の計算コストが比較的高くなる。

連続HMMと比較して計算コストの低い離散HMM＊5を用いた手法が提案されている。各シーンを構成するショットごとにフレーム単位で画像の特徴を計算し，その特徴系列に基づいてショットを７種類のシンボルのいずれかに割り当て，シーンをシンボル列で表現してから分類をする[12]。この手法では，シンボル化のための画像の特徴抽出

＊１　字幕放送。アナウンサーの実況や出演者のせりふを音声認識または人手によってテキストデータ化したもの。

＊２　多変数間の相関を考慮した標本（ベクトル）間の距離。新たな標本と既知の標本との類似性を明らかにするために有用で，統計分類に幅広く使われる。

＊３　システムにある条件（マルコフ過程）を仮定し，観測可能な情報から未知のパラメーターを推定する確率モデル。時間的に連続で伸縮する信号列のパターン抽出に適した確率モデルであり，音声認識，自然言語処理などに広く応用される。

＊４　連続値で表現されたデータを学習および認識対象とする隠れマルコフモデル。

＊５　有限個のシンボルの組み合せで表現されたデータを学習および認識対象とする隠れマルコフモデル。連続HMMと比較して学習時間が短い。

の単位がフレームであり，ショットが持つ画像の特徴系列のバリエーションが広く，多種のプレー種の分類ができると思われるが，実際には，ショットに割り当てられるシンボルの種類が少なく，多種のプレー種の分類は難しい。

提案手法では，シーンの各ショットを推移の基本単位とする。特に，投球の次のショット（以下，第２ショットと呼ぶ）では，ショットを一定の時間間隔で分割した部分ショットを推移の単位とする。部分ショットを推移の単位とすることで，プレー種を区別するための重要な情報を含む映像区間の画の推移やカメラワークのパターン表現を詳細に分析でき，分類精度が向上すると考えた。提案手法では，まず，ショットおよび部分ショットを大域的に簡略化してシンボルで表現し，シーンをシンボル列で表現する。次に，これらのシンボル列の集合をプレー種ごとに学習した離散HMMを用いて，野球放送の映像シーン分類を行う。映像区間の簡略化では，計算コストおよび分類精度に影響するコードブックのシンボル数を抑えながら，より多くのプレー種の分類を可能にしている。更に，イレギュラーな時間軸方向の伸縮や不必要な映像の短時間な挿入がある映像に対しても，精度を低下させることなく，ロバストなプレー種の自動分類を実現している。

２．シーン表現と学習処理

映像の識別単位であるシーンの定義と識別対象のプレー種および離散HMMを用い　た学習処理の詳細について述べる。

2.1　シーンの定義

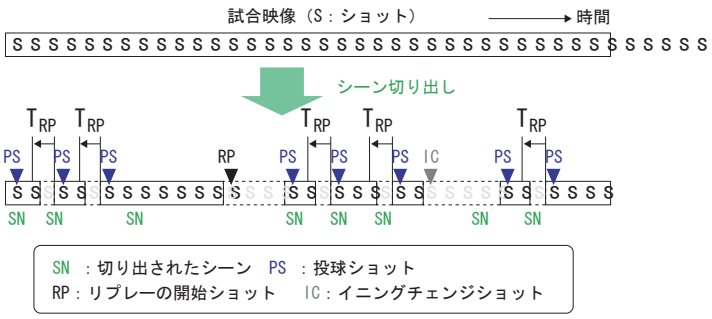
プレー種の識別の単位となるシーンの定義および切り出しについて述べる。シーンは１つ以上のショットから構成され，１回の投球に対応するセグメントであり，先頭ショットは常に投手の投球ショットとする。また，最終ショットは以下の３つの条件のいずれかとする。

- R１：プレーシーン開始の直前のショット。
- R２：イニングの終わりに「スコアボードCG」がスーパーされる「イニングチェンジショット」の直前ショット。
- R３：上記以外の場合で次の投球ショットの開始フレームからT_{RP}フレームさかのぼった時点のショット。

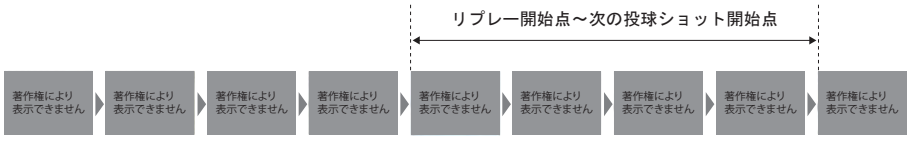
R１およびR２の条件は，中継映像と実時間の異なるリプレー区間や，プレー内容に無関係な映像を多く含むイニングチェンジ区間を除くための条件である。また，R３は投球ショットの間にリプレーシーンがない場合でも，あるときとほぼ同じ方法でシーンを切り出すための条件である。１図に３つの条件を用いたシーンの切り出しの概念図を示す。なお，T_{RP}の値としては，４章の実験で使用する野球映像の一部を用いて測定したリプレーシーン開始から次の投球ショット開始までの長さ（２図の矢印区間）の平均値を採用して463とした。

３つの条件に従ってシーンを自動的に切り出しするためには，ショット分割処理および投球ショット，リプレー開始ショット，イニングチェンジショットの検出が必要である。ショット分割については，映像処理の分野でさまざまな研究が行われており，高速で高精度な検出手法が数多く提案されている[13)~16)]。また，投球ショットの検出も高精度な検出手法が実現されている[17)~19)]。リプレー区間の検出についても盛んに研究されており，良好な精度を示す多くの先行研究がある[20)~23)]。また，イニングチェンジ区間については，次の投球ショットまでのインターバルが長いこと，スコアボードスーパーがほぼ必ず画面の下部に表示されることなどを手がかりとすれば容易に検出可能であると推測した。

そこで，本稿では，ショット検出，投球ショット，リプレー開始ショットおよびイニングチェンジショットの検出は正確に抽出できるとして，議論を進める。



1図　シーンの切り出し



2図　設定のためのリプレー区間

2.2　識別対象プレー種

以下の７種のプレー種を識別対象とした。

- ホームラン
- ツーベース
- シングルヒット
- フライアウト（フライでアウト）
- 内野ゴロ（内野安打を含む）
- 四球
- 三振

上記の７種を採用した理由は，画の推移パターンの傾向が比較的つかみやすいと予想されたからである。３図にシーンの例を示す。シーン１，２がツーベース，シーン３，４がホームランの例である。

2.3　学習処理の流れ

学習用シーンを大域的にとらえてシンボル列化し，プレー種ごとのシンボル列を学習した離散HMMを用いてプレー種の識別処理を行う。学習処理の流れを以下に示す。各処理の詳細は2.4節以降で述べる。

１．学習用の試合映像から，2.2節で述べた７種のいずれかに該当するシーンを2.1節のルールで切り出し，各プレー種へ振り分ける。

２．各プレー種のすべてのシーンにおけるショット転換点を検出する。また，第２ショットを固定長の部分ショットに分割する。

３．ショットおよび部分ショットを画像の特徴に基づいて簡略化した簡略データで表現する。

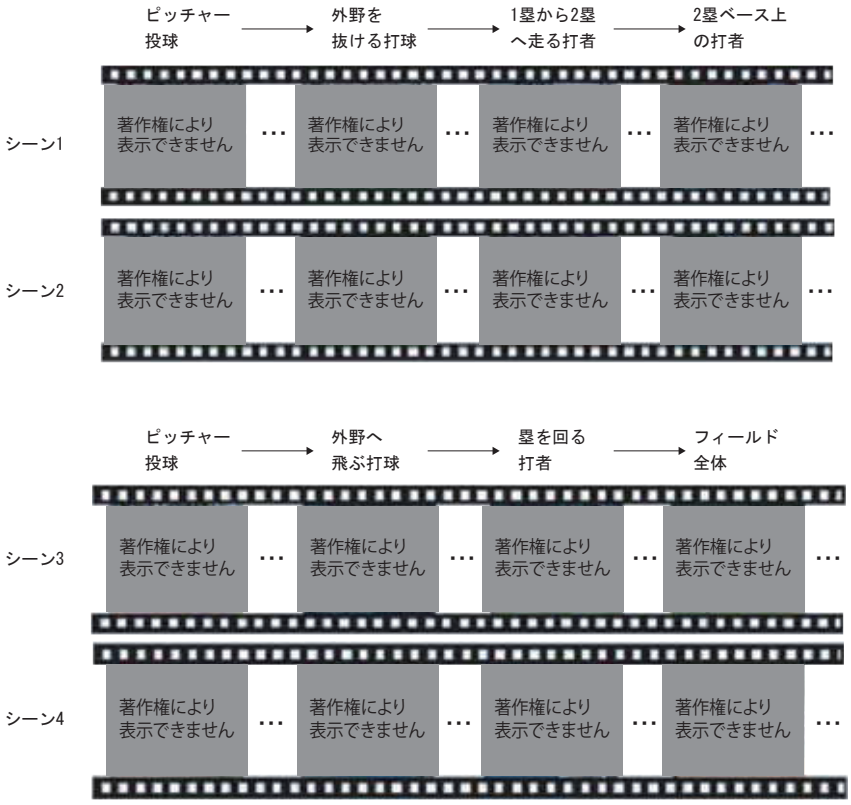
４．簡略データをクラスタリング処理*6し，簡略データとシンボルを対応付けるためのコードブックを作成し，コードブックに基づいて各シーンをシンボル列へ変換する。

５．プレー種ごとにシンボル列集合をHMMに学習させる。

2.4　第２ショットの分割

投球ショットに続く第２ショットは，打球の方向，野手の捕球および送球など，プレー種を区別するための重要な情報を多く含んでいる。そこで，第２ショットに含まれる情報を学習データに強く反映させるために，第２ショットについては，ショットをT_{SS}フレーム間隔で分割した部分ショットを処理単位とする。T_{SS}は「内野ゴロシーンの第２ショット」において，野手が捕球したタイミングで区切ることを主目的としている。野手の送球に

＊６　データの集合を共通の特徴を持つ部分集合（クラスター）へ自動分配する処理。



3図　プレー種ごとの画の推移例

よってカメラの動く方向が大きく変わるので，そのタイミング付辺で分割するのが望ましいと考え（４図），学習用の内野ゴロのシーンを精査して，T_{SS}=45と定めた。第２ショットを分割して部分ショットを生成した例を５図に示す。以下，ショットおよび部分ショットを単位映像区間と呼ぶ。

2.5　単位映像区間の簡略化

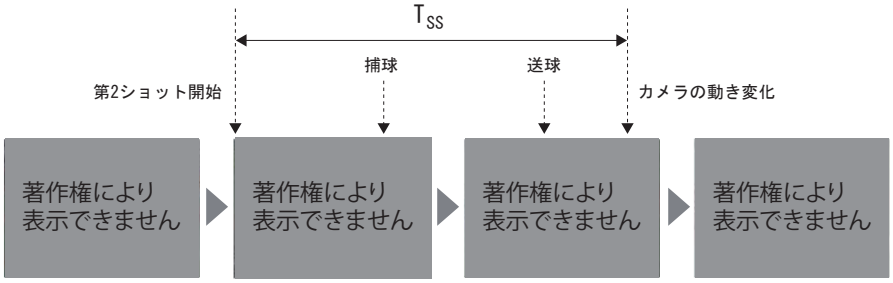
提案手法では，シーンの単位映像区間の並びをシンボル列として表現する。そのために，まず，各単位映像区間の情報を簡略化した簡略データ（以下，SD（Simplified Data））を作成し，シーンをSD列で表現する。SDは後述するように，位置情報，画像情報および動き情報を持つ。簡略化の手法，すなわち，SDを作成する手法としては単位映像区間のHSV（色相，彩度，明度）ヒ ストグラムに基づく手法[24]などが提案されているが，我々が既に提案している手法では，色が同じ領域だけでなく，複雑な画像（テクスチャー）領域，例えば，野球場の観客席やベンチにいる選手の集合なども１つのオブジェクトとして抽出することができる[25][26]。そこで，我々が提案している手法に準じた手法で簡略化することにした。以下，１つの単位映像区間を簡略化する手法の大まかな流れを述べる。

１．単位映像区間にあるすべてのフレームを，例えば，横16×縦12のブロック（小領域）に分割する。ブロックの総数は16×12×（単位映像区間内のフレーム数）である。すべてのブロックを画像特徴ベクトル（HSV平均値，エッジ比率などの画像（テクスチャー）の特徴で構成されている）に基づいてクラスタリングし，テクスチャークラスター群を形成する。１つのテクスチャークラスターに属するブロックはほぼ同じ画像の特徴を持っており，そのテクスチャークラスターの中心にあるブロック，すなわち，そのテクスチャークラスターを代表するブロックの画像を中心画像と呼ぶことにする。

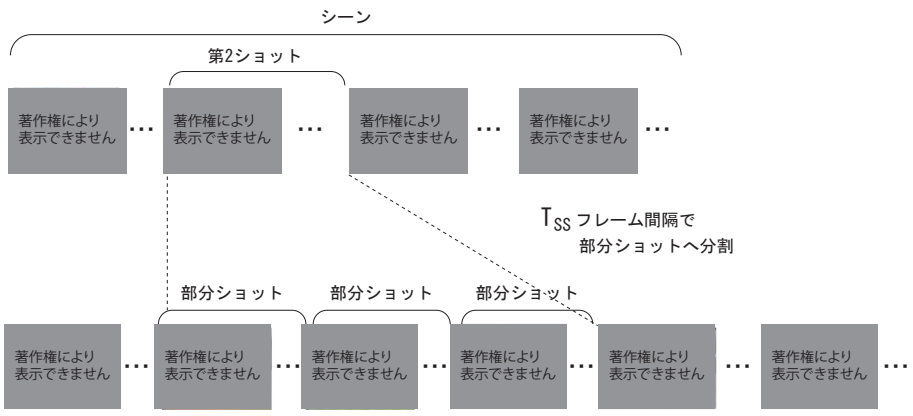
２．単位映像区間の先頭フレームの192（＝16×12）個のブロックに，それぞれ最も類似した中心画像を割り当てる。

３．先頭フレームにおいて，同じ中心画像が割り当てられた連続したブロックをグループとしてまとめる（ブロックグループ）。なお，グループ内のブロック数があらかじめ設定した設定値よりも小さいグループはノイズとしてブロックごと除去する。

４．３の処理で残ったブロックを対象として，先頭フレームから最終フレームまでブロックマッチング処理を行い，それぞれのブロックが先頭フレームから最終



4図　T_{SS} 設定の目安



5図　第2ショットの分割

フレームまでどのように移動するのかを追跡した動きベクトルを算出する。

5．３の処理で得られるブロックグループの形は矩形または幾つかの矩形を組み合わせた形になる。矩形ではないブロックグループでは必要最小限のブロックを補充して，すべてのブロックグループの形を矩形にする。

6．以下の要素を持つ矩形構造体を定義する。

- 位置情報：５の処理で得られた矩形の４隅の座標
- 画像特徴：２の処理で割り当てた中心画像の特徴ベクトル
- 動き情報：ブロックグループごとに４の処理で得られたブロックごとの動きベクトルの方向を６等分に量子化した（６図（c）参照）ヒストグラム。

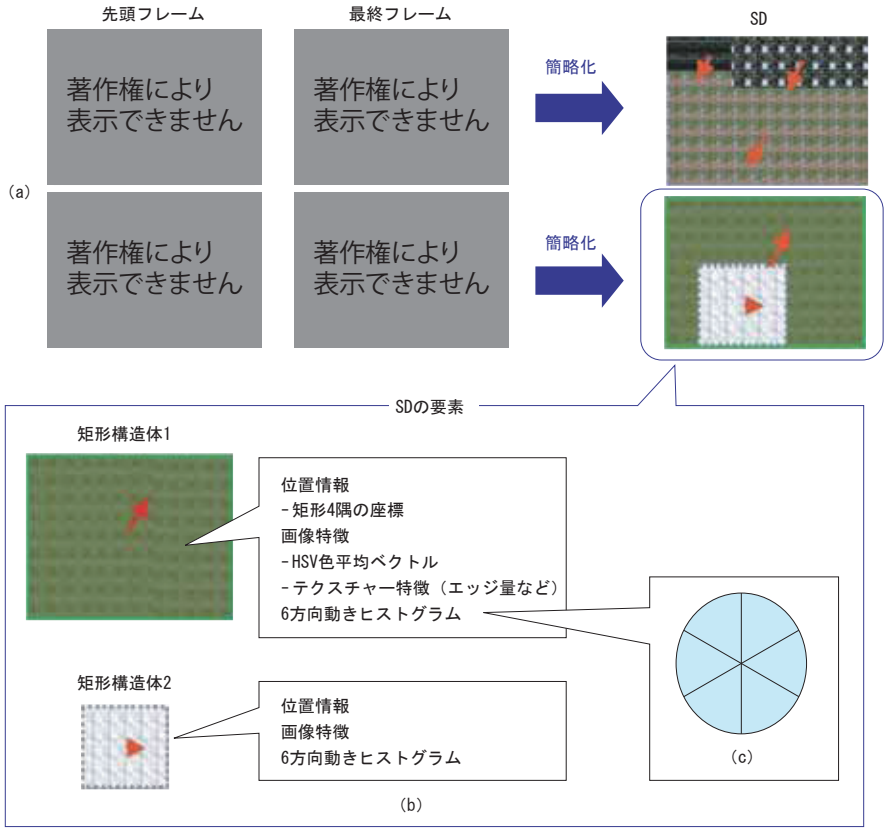
7．３の処理で得られたブロックグループごとに矩形構造体を作成し，SDを複数の矩形構造体で表現する。SDは単位映 像区間ごとに生成され，シーンのシンボル列化（2.6節）で用いられる基本的な特徴量である。

６図（a）にSDを視覚的に表現した例を示す。６図（a）のSDにおいて，同じ中心画像で敷き詰められた領域が１つの矩形構造体である。また，矢印は動きヒストグラム中でもっとも高い数値の長さとその方向を表したものである。６図（a）では，大きな矩形構造体の上に小さな矩形構造体を重ねて表示したので，６図（b）に示す矩形構造体１が矩形には見えていない。

2.6　簡略データ列のシンボル列への変換

SD列で表現さ れたシーンをシンボル列へ変換するためには，SDとシン ボルを対応付けるコードブックが必要である。まず，学習用シーンにおけるSDの集合をシーン中での出現位置に基づいて，以下の４つの時間的SDグループ（以下，TSDG（T　iming SD Group）と呼ぶ）に分類する。ここで，投球ショット直後の先頭の部分ショット（第２ショットの１番目の部分ショット）は打球の方向に関する情報を多く含むので，それ以降の部分ショットとは区別して扱うこととした。

- TSDG−A：投球ショットに対応したSDのグループ
- TSDG−B：先頭の部分ショットに対応したSDのグループ
- TSDG−C：２番目以降の部分ショットに対応したSDのグループ
- TSDG−D：第３ショット以降のショット（後半ショット）に対応したSDのグループ



6図　SDを視覚的に表現した例((a)の右側2枚)　とSDの詳細

コードブックをTSDGごとに作成した。これは，異なるTSDGでは，同じようなショットでも異なるシンボルに変換でき，各シーンにおけるシンボルの推移をより正確に表現できるので，プレー種の分類の精度が向上すると考えたからである。**7図**にコードブックを作成するための全体の流れを示す。学習用シーンのすべてのSDを各TSDGに振り分け，TSDGごとにSDの階層的クラスタリング処理*7を行い，コードブックを作成した。各TSDGでコードブックを作成するための処理の概念図を**8図**に示す。処理の流れは以下のとおりである。

1　1つのTSDGに属するすべてのSDに対応する先頭フレームの画像をそれぞれ4×4程度の小領域に分割する。小領域ごとの画像の平均的な特徴ベクトルのユークリッド距離に基づいて類似度Sim 1*8を計算し，Sim 1に基づいてクラスタリングを行う。

2　1の処理で生成したすべてのクラスターについて，SDにおいて面積が最大の矩形構造体（ほとんどの場合は背景領域に該当する）の動きヒストグラム類似度Sim 2を計算し，Sim 2に基づいたクラスタリングを行う。

3　2の処理で生成したすべてのクラスターについて，上部および下部（観客席や地面の可能性がある）を除いた中央付近に存在する矩形構造体（人物の領域に該当する可能性が高い）の矩形面積の類似度Sim 3を計算し，Sim 3に基づいたクラスタリングを行う。

4　3の処理で生成されたすべてのクラスターにシンボル（数値ラベル）を割り当てる。
生成したTSDGごとのコードブックを用いて各SDをシンボルに変換し，すべての学習用シーンをシンボル列で表現する。

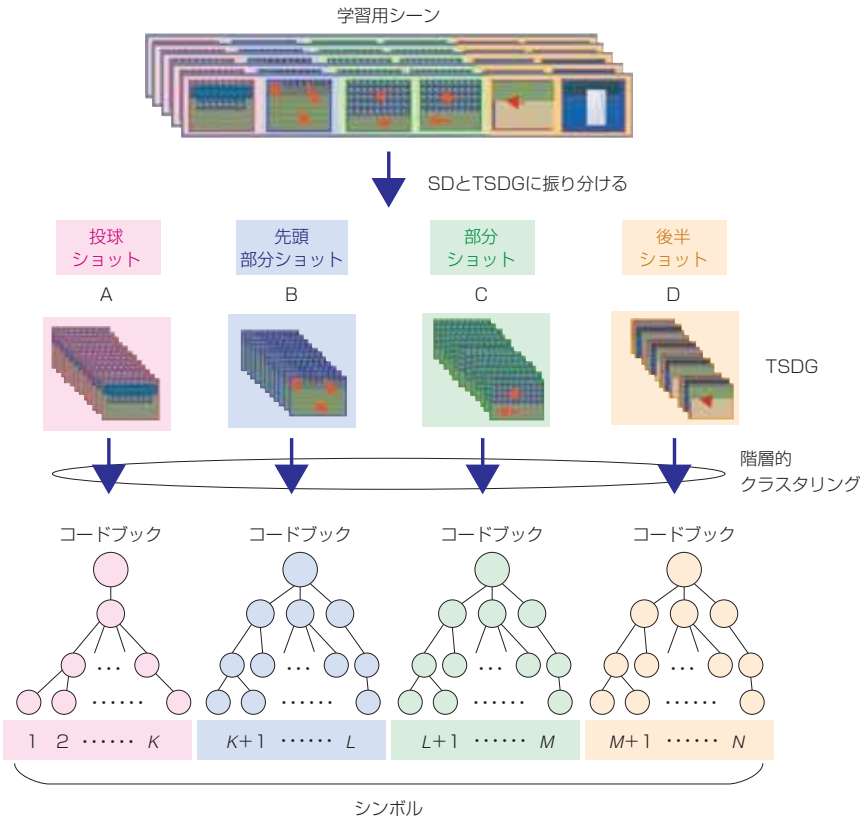
コードブックを作成するための1～3の処理はそれぞれ，絵柄の全体的な類似度，カメラの動きの類似度，人物のアップの度合いの類似度に基づく分類を意図したものである。SDは映像の簡略表現なので，人物が小さく撮影されている場合にはそれが矩形構造体として現れないケースもあるが，1の処理で，人物領域の有無やその位置をある程度反映させることが可能である。

2.7　離散HMMによるシンボル列の学習

2.6節で生成した学習用のシンボル列の集合をプレー種

＊7　クラスタリング処理前のデータ集合を親クラスター，クラスタリング処理で生成したクラスター群を子クラスターと呼び，各子クラスターを親クラスターとしてあらかじめ設定した回数だけクラスタリングを繰り返す処理。

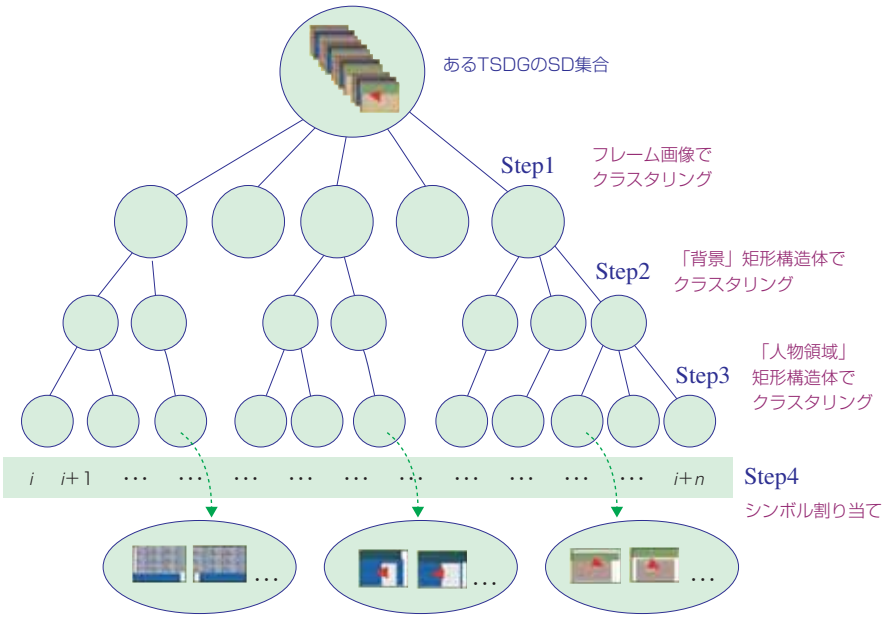
＊8　色やテクスチャーなどを比較する一般的な画像類似度。



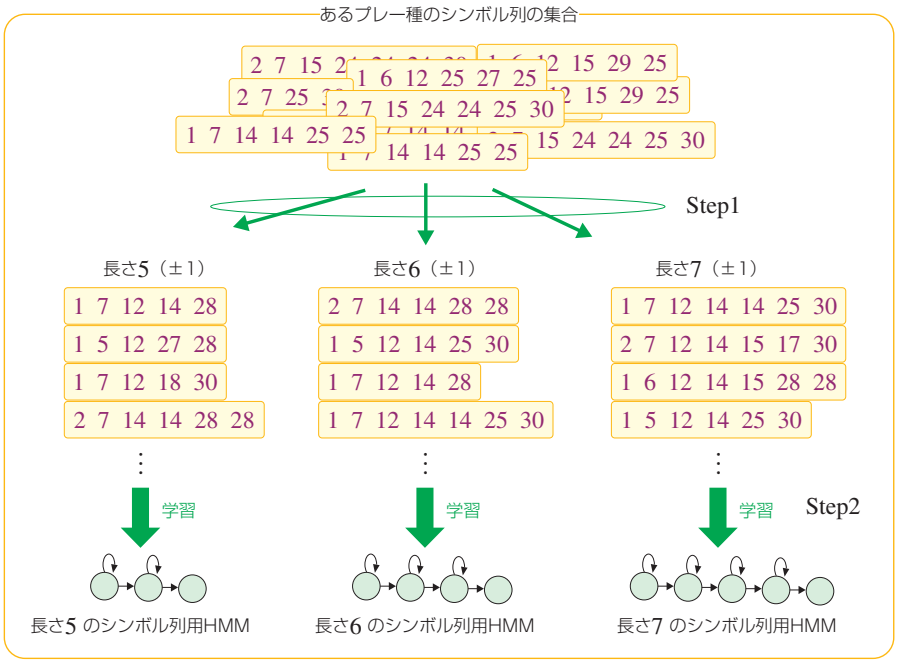
7図　コードブック作成の全体の流れ

ごとに準備した複数の離散HMMに学習させる。9図にHMMによるプレー種ごとの学習処理の流れを示す。

Step 1では，プレー種に属するシンボル列を長さの近いシンボル列のグループに分類する。具体的には，長さ$L-1$，L，$L+1$のシンボル列を束ねて「長さLのシンボル列グループ」とする。シンボル列の長さはシーンを構成するショット数を反映しており，プレー種の分類の重要な手がかりとなる（例えば，ホームランシーンは構成ショット数が多く，三振や四球は少ない傾向がある）と考えられる。分類処理において入力シーンと近い長さのシンボル列を学習したHMMのみを用いることで，可能性の著しく低いプレー種の尤度計算を回避することができる。Step 2では，それぞれの長さのシンボル列グループに対して，個別に離散HMMを準備し，グループの属するシンボル列を



8図　各TSDGにおける階層的クラスタリング処理によるコードブック作成



9図　HMMの構築

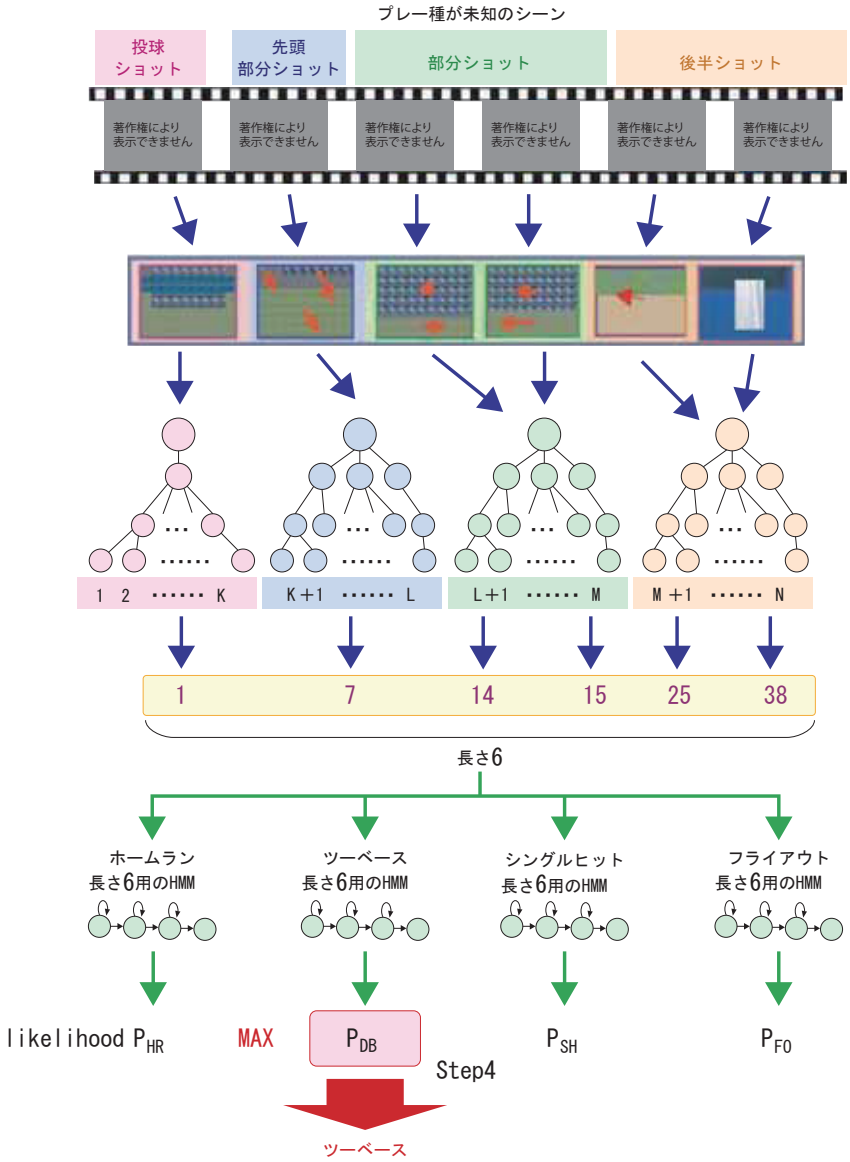
学習させてパラメーターを決定する。ここで使用するHMMのモデルは，最　も一般的なlefttorightモデル *9で，状態数はシンボル列長の2/3（端数切り上げ）とした。離散HMMを用いた処理は　，数十秒から数分といった非常に低い計算コストで実行可能である。

3．プレー種分類処理

プレー種が未知のシーンに対するプレー種識別処理は，各プレー種において構築されたHMMを用いて行う。　10図に分類処理の流れを示す。まず，入力シーンをショット分割し，学習データ生成時と同様に，単位映像区間ごとに簡略化の処理をしてシンボル列化する。各SDのシンボル変換には，学習時に生成したコードブックを用いる。具体的には，対応するTSDGのコード　ブックの上層から下層へ，クラスター中心（平均）データとの類似度（2.6節で用いた類似度で，上層から順にSim１，Sim２，Sim３）の最も高い子クラスターへたどっていき，葉クラスター*10に割り当てられているシンボルへ変換する。各プレー種におけるHMM群の中からそのシン　ボル列と同じ長さのシンボル列グループを学習したものだけを用いて認識処理を行い，それぞれのプレー種である確からしさ（尤度）を算出する。その尤度が最大となるプレー種をこの未知のシーンのプレー種とする。

＊９　一定方向だけに状態が遷移する隠れマルコフモデル。

＊10　階層的クラスタリングの繰り返し処理において，最後のクラスタリング処理で生成される子クラスター。



10図　プレー種の分類処理の流れ

４．プレー種分類実験

18試合分の映像ファイルを３つのセットに分け，１セットを実験用，２セットを学習用として相互検証実験を行った。各セットの映像ファイルの詳細データを１表に示す。１表に示すように，映像ファイルは球場，時間帯，現地での放送チャンネルの各条件が同じものだけではなく，幾つかのバリエーションがある。

2.5節で述べた簡略化の処理のメリットが本実験において生かされていることを11図の例によって示す。11図（a）の上段と下段は非常に似た映像であるが，上段には「人の横切り」がある。しかし，それには影響されずに上下段共にほぼ類似したパターンデータが出力されており，同じシンボルの割り当てが実現できている。また，上段と下段で「人の横切り」の回数が異なる11図（b）においても，同様に同じシンボルが割り当てられている。11図（c）の上段は投手がいったんプレートを外したので，下段と比較して著しく長い映像になっているが，上下段共に同じシンボルが割り当てられている。

２表に提案手法によるプレー種の分類結果を示す。実験１ではSet１を実験用に，Set２，３を学習用に使用し，実験２ではSet２を実験用に，Set１，３を学習用に使用し，実験３ではSet３を実験用に，Set１，２を学習用に使用した。実験では，すべてのシーンが必ず７つのプレー種のいずれかに分類されるので，合計欄の再現率と適合率

1表　ショット境界の検出精度

(a) Set1

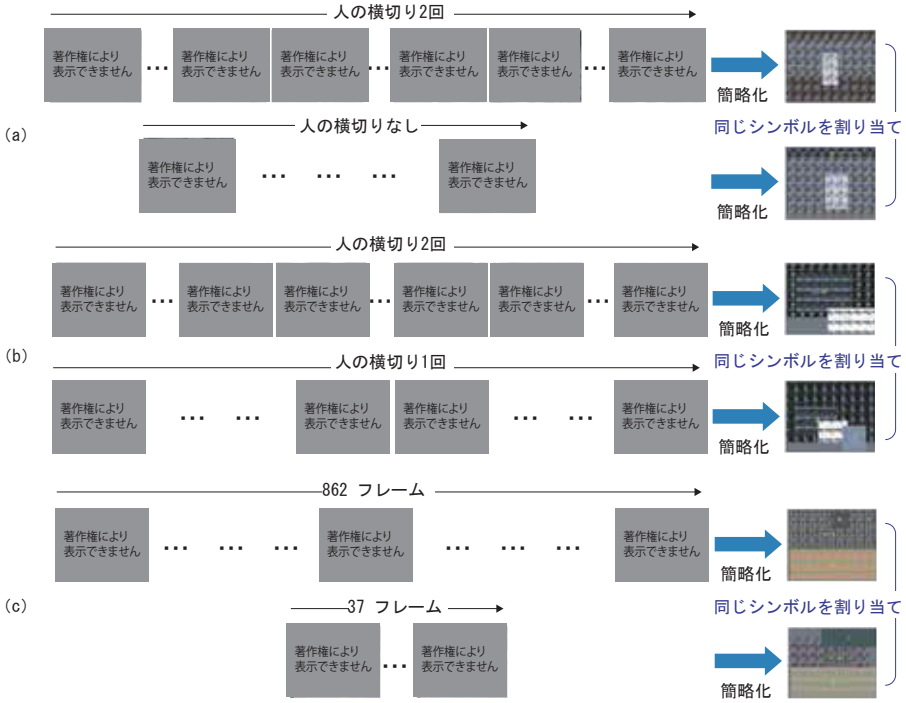
ファイル	対戦カード	球場ラベル	夜(N)/昼(D)	放送Ch
V01	SEAOAK	1	N	A
V02	SEAANA	3	N	A
V03	MINNYY	4	N	B
V04	NYYTRB	5	D	C
V05	ANASEA	2	N	B
V06	MINSEA	2	D	B

(b) Set2

ファイル	対戦カード	球場ラベル	夜(N)/昼(D)	放送Ch
V07	SEAOAK	1	N	A
V08	ANASEA	2	N	B
V09	MINSEA	2	N	B
V10	NYYTRB	4	N	C
V11	BALNYY	4	D	B
V12	BALNYY	4	D	B

(c) Set3

ファイル	対戦カード	球場ラベル	夜(N)/昼(D)	放送Ch
V13	MINSEA	2	D	B
V14	NYYSEA	2	N	B
V15	NYYSEA	2	D	B
V16	BALNYY	4	D	B
V17	NYYTRB	5	D	C
V18	NYYBOS	6	D	A



11図　ノイズに影響されずにシンボルが割り当てられている例

2表　提案手法による実験結果

	実験1			実験2			実験3		
プレー種	実験シーン数	再現率(%)	適合率(%)	実験シーン数	再現率(%)	適合率(%)	実験シーン数	再現率(%)	適合率(%)
ホームラン	13	69.2	64.3	13	69.2	90.0	13	84.6	55.0
ツーベース	12	33.3	44.4	3	33.3	12.5	16	6.2	33.3
シングルヒット	48	66.7	64.0	55	58.2	72.7	48	68.8	50.8
四球	28	46.4	48.1	38	55.3	46.7	25	56.0	43.8
フライアウト	96	69.8	71.3	88	71.6	70.0	111	62.2	73.4
内野ゴロ	122	96.7	95.2	85	94.1	87.0	81	90.1	83.9
三振	61	77.0	75.8	57	57.9	66.0	73	75.3	83.3
計／平均	380	76.3	76.3	339	70.5	70.5	367	69.8	69.8

3表　映像ファイルごとの全実験の中で最も精度の高かった実験結果

映像ファイルV04（球場ラベル：5）

プレー種	実験シーン数	再現率(%)	適合率(%)
ホームラン	1	100.0	100.0
ツーベース	0	–	–
シングルヒット	10	90.0	75.0
四球	6	33.3	40.0
フライアウト	13	76.9	90.9
内野ゴロ	29	100.0	100.0
三振	7	57.1	50.0
計／平均	66	83.3	83.3

は同じ値になり，約70~76％の精度であった。

3表は18試合分の実験用映像ファイルの中で，最も高い分類精度を得た映像ファイルV04での実験結果を示したものである。1表を見ると，この実験における学習用映像ファイル（Set 2，3）の中に，V04と同じ球場で行われたものが少ないことがわかる。それにもかかわらず高い精度が得られたことは，提案手法は，すべての条件での野球映像を網羅するような大量の学習データを必要とせず，数十試合程度の扱いやすい学習データ量で高精度のプレー種の分類ができることを示している。

従来手法の中には，提案手法の平均精度よりも高い分類精度を得ている手法[11]があるが，1組のデータセットによる結果のみが示されており，相互検証実験を行っていないなど，検証が不十分である。我々は，より多くのデータセットで相互検証実験を行い，一部の映像ファイルについては，従来手法を上回る精度を得た。提案手法は信頼度の高いロバストなプレー種の分類を実現していると言える。

ただし，プレー種ごとの精度を評価すると，ツーベースおよび四球において，十分な分類精度が得られていない。ツーベースについては，前述の従来手法[11]においても分類精度が低く，学習データが不足していることが原因であると考えている。また，四球については，打者が1塁へ移動する四球シーンと打者が1塁ベンチへ帰る三振シーンの先頭ショット前半のカメラワークが酷似していることが主な原因と考えている。先頭ショットを第2ショットと同様に部分ショットへ分割することで更なる改善が期待できる。

第2ショットの分割による効果を検証するために，比較対象として第2ショットを分割しないで同様の実験を行った。4表に結果を示す。2表と比較して明らかなように，いずれの実験においても，主に打球が飛ぶイベント（三振，四球以外）で提案手法の効果が現れており，精度が約7～9％向上している。

5．むすび

第2ショットを分割することで，プレー種を区別するための重要な情報（打球の方向，野手の捕球および送球など）を含む映像区間の画の推移やカメラワークのパターンを詳細に表現することができ，分類精度が向上すると考え，第2ショットに重きを置いたシーンのシンボル列表現と離散HMMを用いて，野球映像の各シーンを自動的にプレー種分類する新しい手法を提案した。提案手法では，シーン中の各ショットおよび第2ショットを分割した部分ショットを簡略化し，それをシンボルに変換する。コードブックのシンボル数を抑えることができ，少ない学習データで高い分類性能が得られる。更に，長い投球間隔，人が横切るあるいはミススイッチングなどの「揺らぎ」に対してロバストなイベント分類が可能である。

提案手法は，蓄積された大量の野球アーカイブス映像に自動的にメタデータを付与し，素早く目的のシーンを探し出すことを可能とする。また，野球に強い興味を持つユーザーにとって有用なWEBサービス（チームや打者の打撃傾向，投手の球種ごとの被打率などの「解析データ」の配

4表　第2ショットを分割せずに行った実験結果

	実験1			実験2			実験3		
プレー種	実験シーン数	再現率(%)	適合率(%)	実験シーン数	再現率(%)	適合率(%)	実験シーン数	再現率(%)	適合率(%)
ホームラン	13	46.2	60.0	13	61.5	61.5	13	53.8	58.3
ツーベース	12	8.3	33.3	3	33.3	14.3	16	0.0	0.0
シングルヒット	48	58.3	54.9	55	27.3	38.5	48	50.0	32.4
四球	28	71.4	47.6	38	65.8	59.5	25	60.0	36.6
フライアウト	96	65.6	62.4	88	52.3	63.0	111	64.0	73.2
内野ゴロ	122	84.4	81.7	85	88.2	67.0	81	84.0	81.9
三振	61	63.9	83.0	57	70.2	75.5	73	64.4	82.5
計／平均	380	68.4	68.4	339	61.9	61.9	367	63.2	63.2

信）実現への貢献も考えられる。

今後の課題としては，まず，学習データを増やす必要がある。提案手法は大量の学習データを必要とはしないが，常に安定した分類結果を得るためには，さまざまな状況の変化に対してロバストな学習データを集める必要があり，ある程度の学習データの増加は不可欠である。また，簡略データの精度や，コードブック作成時の簡略データのクラスタリング精度も学習データの質を上げるためには非常に重要な要素であり，今後，改良していく必要がある。

本稿では，出塁やアウトカウントの増加が生じる「選手の打席が完了する」７つのプレー種の分類を行ったが，試合映像の中には，投球のみ，ファウル，けん制，盗塁など，「選手の打席が完了しない」シーンも含まれており，これらの識別も可能とする改良が必要であると考えている。

本稿は，電子情報通信学会論文誌に掲載された以下の論文を元に加筆，修正したものである。

望月，藤井，八木，篠田：”投球の次ショットに重きを置いたシーンのパターン化と離散隠れマルコフモデルを用いた野球放送映像の自動イベント分類，”映像情報メディア学会誌，Vol.61 No.8，pp.1139−1149（2007）

参考文献

1） NHKアーカイブス，http://www.nhk.or.jp/archives/

2） NHKオンデマンド，https://www.nhk-ondemand.jp/

3） NHKクリエイティブライブラリ，http://cgi4.nhk.or.jp/creative/cgi/page/Top.cgi

4） N. Dimitrova, H. J. Zhang, B. Shahraray, I. Sezan, T. Huang and A. Zakhor：“Applications of Video-Content Analysis and Retrieval,”IEEE Multimedia, 7（3），pp.42-55（2002）

5） N. Babaguchi, Y. Kawai and T. Kitahashi：“Event Based Video Indexing by Intermodal Collaboration,”Proc. of First Int. Workshop on Multimedia Intelligent Storage and Retrieval Management, pp.1-9（1999）

6） A. Watanabe, K. Komiya, J. Usuki, K. Suzuki and H. Ikeda：“Effective Designation of Specific Shots on Video Service System Utilizing Mahalanobis Distance,”IEEE Trans. Consum. Electron., 51, 1, pp.152-159（2005）

7） A. Ekin, A. M. Tekalp, and R. Mehrotra：“Automatic Soccer Video Analysis and Summarization,”IEEE Trans. Image Processing, 12, 7, pp.796-807（2003）

8） G. Xu, Y. F. Ma, H. J. Zhang and S. Yang：“Motion Based Event Recognition Using HMM,”Proc. of the Int. Conf. on Pattern Recognition, pp.831-834（2002）

9） L. Xie, S. F. Chang, A. Divakaran, and H. Sun：“Structure Analysis of Soccer Video with Hidden Markov Models,”Proc. ICASSP 2002, 4, pp.13-17（2002）

10） N. Rea, R. Dahyot and A. Kokaram：“Classification and Representation of Semantic Content in Broadcast Tennis Videos,”Proc. of the Int. Conf. on Image Processing, pp.III-1204-1207（2005）

11） H. B. Nguyen, K. Shinoda and S. Furui：“Robust highlight extraction using multi-stream Hidden Markov Models for baseball video,”Proc. of the Int. Conf. on Image Processing, pp.III-173-176（2005）

12） P. Chang, M. Han and Y. Gong：“Extract highlights from baseball game video with hidden markov models,”Proc. of the Int. Conf. on Image Processing, 1, pp.609-612（2002）

13） J. S. Boreczsky and L. A. Rowe：“Comparison of video shot boundary detection techniques,”Proc. SPIE, Vol.2664, pp.170-179（1996）

14） R. Lienhart：“Comparison of automated shot boundary detection algorithms,”Proc. SPIE, Vol.3656, pp.290-300（1999）

15） Y. Kawai, H. Sumiyoshi and N. Yagi：“Fast detection method for shot boundary including gradual transition using multiple features,”IEICE Tech. Rep., CS2007-53, pp.141-146（2007）

16） 望月，蓼沼，八木：“静止画と音声によるドラマ番組配信のための要約画像系列生成,”信学技報，IE2005-21, pp.19-24（2005）

17） W. N. Lie, G. S. Lin and S. L. Cheng：“Pitching Shot Detection Based on Multiple Feature Analysis and Fuzzy Classification,”Proc. of Pacific-Rim. Conf. on Multimedia, pp.852-860（2006）

18） Y. Ariki, M. Kumano and K. Tsukada：“Highlight Scene Extraction in Real Time from Baseball Live Video,”Proc. of ACM Multimedia Int. Workshops, Multimeida Information Retrieval, pp.209-214（2003）

19） M. H. Hung, C. H. Hsieh and Y. C. Zhu：“Scene Classification for Baseball Videos Using Spatial and Temporal Features,”9th JCIS 2006, pp.1053-1056（2006）

20） 河合，馬場口，北橋：“放送型スポーツ映像におけるディジタルビデオ効果に着目したリプレイシーン検出の一手法,”信学論，Vol. J84-D-II, No. 2, pp.432-435（2001）

21） N. Babaguchi, Y. Kawai, Y. Yasugi and T. Kitahashi：“Linking Live and Replay Scenes in Broadcasted Sports Video,”Proc. of ACM Multimedia Int. Workshops, Multimeida Information Retrieval, pp.205-208（2000）

22）H. Pan, B. Li, and M. Sezan：“Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transition,” Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.3385–3388（2002）

23）H. Pan, P. Beek, and M. Sezan：“Detection of slow–motion replay segments in sports video for highlights generation,” Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp.1649–1652（2001）

24）T. Lin and H. J. Zhang：“Automatic Video Scene Extraction by Shot Grouping,” Proc. of the Int. Conf. on Pattern Recognition, 4, pp.39–42（2000）

25）望月，藤井：“映像検索のための映像の簡易表現の一手法,”信学総大講演論文集，情報・システム（2），p.280（2003）

26）T. Mochizuki, M. Tadenuma and N. Yagi：“Baseball video indexing using patternization of scenes and hidden Markov model,” Proc. IEEE Int. Conf. Image Processing,Vol.3, pp.1212–1215（2005）

望月貴裕（もちづきたかひろ）

1996年入局。放送技術局報道技術センター中継制作部を経て，1998年から放送技術研究所にて，画像および映像解析の研究に従事。現在，放送技術研究所人間・情報科学研究部専任研究員。博士（工学）。

藤井真人（ふじいまひと）

1983年入局。札幌放送局を経て，現在，放送技術研究所人間・情報科学研究部主任研究員。この間，CMUに半年間滞在およびATR人間情報通信研究所に出向。神経回路モデル，視覚情報処理，画像認識，映像検索などの研究開発に従事。博士（情報科学）。

八木伸行（やぎのぶゆき）

1980年入局。甲府放送局，放送技術研究所，技術局，編成局を経て，現在，放送技術研究所研究企画部部長。画像・映像・メディア情報処理，コンピューターアーキテクチャー，コンテンツ制作技術，デジタル放送などの研究開発に従事。博士（工学）。

篠田浩一（しのだこういち）

1987年，東京大学理学部物理学科卒業。1989年，同大大学院修士課程終了。同年日本電気（株）入社。以後，音声認識の研究に従事。その間，1997年から1998年にかけて，米国ルーセントテクノロジ・ベル研究所客員研究員。2001年から東京大学情報理工学系研究科助教授，2003年から東京工業大学情報理工学研究科准教授。研究分野は，音声・動画像などに対する統計的パターン認識。博士（工学）。

報告

電子番組表における紹介テキストを利用した番組紹介映像の自動生成

河合吉彦　住吉英樹　八木伸行

Automated Production of TV Program Trailer Using an Introductory Text from Electronic Program Guides

Yoshihiko KAWAI, Hideki SUMIYOSHI and Nobuyuki YAGI

要約

映像要約とは元の映像からより短い映像を生成する操作と定義される。映像要約は大量の映像アーカイブを効率的に検索するための有効な技術の 1 つである。本稿では，要約映像の 1 種である番組紹介映像の自動生成手法を提案する。番組紹介映像の目的は番組の内容を紹介することであり，電子番組表に記載されている番組紹介テキストの目的と同じである。提案手法では，番組紹介テキストの各文に最も類似しているクローズドキャプションを探索し，そのクローズドキャプションに対応する映像区間を抽出して連結する。番組紹介テキストとクローズドキャプションの類似度の算出には，ベイズ信頼度ネットワークを利用する。実際の放送映像に提案手法を適用し，複数の専門家が作成した正解データと映像内容を比較した結果，番組紹介テキストを利用しないでクローズドキャプションだけを利用する従来手法よりも良好な結果が得られた。

ABSTRACT

Video abstraction is defined as producing shorter video clips from the original video and it is one of the most efficient methods for retrieval from large video archives. We propose an automated method of producing TV program trailers. Our method employs introductory text from an electronic program guide, which is a short description of the program highlights. We extract closed caption sentences that have the highest similarity for each introductory sentence and then connect the corresponding video segments to make the trailer. A Bayesian belief network is used to calculate the similarity. The proposed method was used to generate trailers for actual TV programs, by which their effectiveness was verified.

1．まえがき

映像要約とは元の映像から必要なシーンを抽出し，連結することによって，より短い映像を生成する操作と定義される[1]。大量に蓄積された放送映像の中から，所望の番組を効率的に検索したり，番組の内容を短時間で把握したりするための有効な技術の１つである。映像要約は制作目的の違いによってサマリー（summary）型とトレーラー（trailer）型に大別することができる。サマリー型の要約映像は元の映像の代替となることを目的としている。得点シーンなどを集めたスポーツニュースの野球のダイジェスト映像などがこれにあたる。ユーザーがサマリー型の要約映像を視聴することによって，元の映像の全体を理解できることが求められている。これに対して，トレーラー型の要約映像は番組本編を視聴していないユーザーに番組の内容を紹介することを目的としている。映画の予告映像やテレビ放送における番組スポット（以下，番組紹介映像と呼ぶ）などはトレーラー型の代表例である。トレーラー型において重要なことは，映像全体をぜひ視聴したいと思わせる演出やストーリーである。

サマリー型の映像要約の研究としては，映画やドラマ，ドキュメンタリー，スポーツなどを対象に多くの手法が提案されている。例えば，出演者のアップや銃声音[2]，ショット切り替えの頻度やBGM（Back Ground Music）[3]などの特徴に基づいて重要シーンを決定し，映画やドラマを要約する手法がある。また，トーク番組やドキュメンタリーを対象として，クローズドキャプション（CC：Closed Caption）＊1から話題転換を表す言い回しを探索する手法[4]や，音声認識を行って単語の出現頻度や共起＊2に基づいて重要シーンを決定する手法[5]が提案されている。更に，スポーツを対象として，打球音やアナウンサーの興奮した音声の特徴[6]や，放送番組におけるリプレーシーン[7]，CCと画像特徴の統合的な分析[8]などに基づいて得点シーンなどの重要イベントを抽出し，試合映像を要約する手法もある。

一方，トレーラー型の映像要約の研究としては，色ヒストグラムやカメラの動き，音情報に基づいてアクション映画の予告映像を生成する手法[9]や，ショット切り替えの頻度やフレーム差分に基づいてアニメ映画の予告映像を生成する手法[10]などが提案されている。映画の予告映像は爆発シーンなどの視聴覚的に派手なシーンを用いて視聴者の印象に強く残るように制作される傾向がある。そのため，従来手法においても視聴覚的な特徴に基づいて映像区間の抽出を試みている。

本研究では，トレーラー型の要約映像の１つである番組紹介映像の生成を目的とする。番組紹介映像はテレビ放送において利用される短い予告映像であり，番組内容を視聴者に宣伝する目的で制作される。番組紹介映像を自動生成するためには，カメラの動きなどの表層的な特徴ではなく，番組内容やテーマなどの意味内容に基づいてシーンを選択する必要がある。そこで，電子番組表（EPG：Electronic Program Guide）に記載されている番組紹介テキスト（以下，EPGテキストと呼ぶ）を利用した要約手法を提案する。EPGテキストは番組視聴前の視聴者を対象にしたテキスト情報で，番組のテーマや内容などが１文から10文程度で簡潔に記述されている。EPGテキストの目的は番組紹介映像の目的と同じなので，提案手法では，EPGテキストに対応する映像区間を抽出して連結することによって，意味内容に基づく番組紹介映像の生成を試みる。EPGテキストに対応する映像区間はEPGテキストと類似したCC文をCCテキストの中から探索して決定する。なお，文の類似度はベイズ信頼度ネットワーク＊3を利用して算出する。

まず，２章で番組紹介映像の生成手法について詳細に説明する。次に，３章で実際の放送番組に対する実験結果を示し，考察する。最後に，４章で全体のまとめと課題について述べる。

2．番組紹介映像の自動生成手法

2.1　手法の概要

１図に提案手法の概要を示す。まず，EPGテキストの各文（以下，EPG文と呼ぶ）に対して，CCテキストの中から最も類似度の高いCC文を探索する。CC文には映像と同期して表示するための時刻情報が付いているので，抽出されたCC文の時刻情報を基にして映像区間を抽出する。この方法では，カメラの切り替え点を考慮せずに映像区間を抽出するので，１つのCC文に対して複数のショット（１台のカメラで連続的に撮影された映像区間）が選択されることがある。最後に，抽出した映像区間を連結して番組紹介映像を生成する。

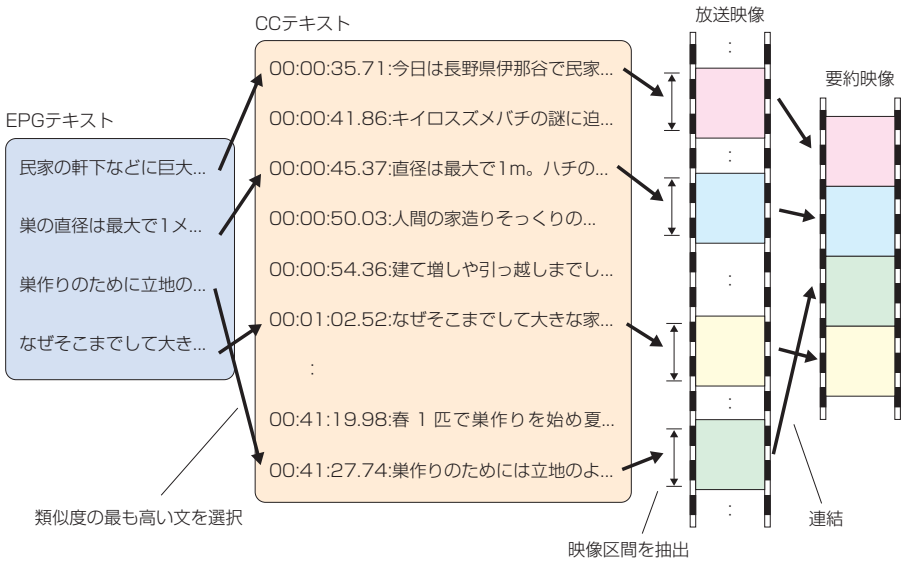
2.2　文の類似性判定に基づく紹介映像生成

EPG文とCC文との対応付けに使用する類似度は２つの文に共通に含まれる単語に基づいて算出する。このとき，助詞や助動詞などの一般的な単語よりも，人名や地名などの固有名詞が共通に多く含まれるほど類似度が高くなるようにする。また，番組の一部の区間だけに出現する単語が

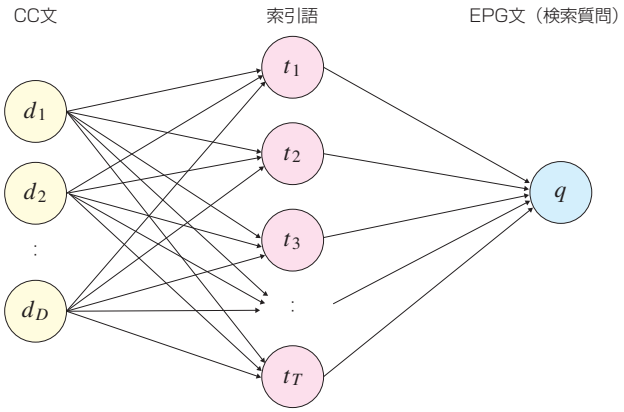
＊１　字幕放送用のテキスト。番組音声を書き起こしたもの。

＊２　複数の単語が同時に出現すること。

＊３　確率的な因果関係を有向グラフでモデル化し，データ間の関係を推論する手法。



1図　提案手法の概要



2図　文の類似度を算出するベイズ信頼度ネットワーク

共通に含まれる文ほど類似度を高くする。提案手法では，このような関係を容易にモデル化できるベイズ信頼度ネットワーク[11]を利用する。

文の類似度を算出するベイズ信頼度ネットワークの構成を**2図**に示す[12]。ノードd_i（$i=1\cdots D$）はCC文に対応し，ノードt_j（$j=1\cdots T$）は単語（以下，索引語と呼ぶ）に対応する。索引語は形態素＊4と品詞の組で表される。qはEPG文である。提案手法では，CC文d_iを与えた場合にEPG文qが成り立つ確率を類似度と定義する。

qが成り立つ確率は（1）式で与えられる。

$$P(q\,|\,d_i)=\sum_{all\ s_k}P(q\,|\,t_{1s_1},\cdots,t_{Ts_T})\prod_{j=1}^{T}P(t_{js_j}\,|\,d_i) \qquad (1)$$

＊4　テキストにおいて意味を持つ最小の単位。

ここで，t_{js_j}は索引語t_jの状態がs_j（$s_j=0$または1）であるという命題を表し，状態0は索引語t_jが存在しないという命題を，状態1は存在するという命題を表す。Tはノードt_jの総数である。なお，（1）式では，直接結合していないノードd_i間およびノードt_j間は統計的に独立であると仮定している。また，ノードd_iは同時に複数選択されないという制限をしている。つまり，EPG文と複数のCC文との類似度については考慮しない。

（1）式における$P(t_{js_j}\,|\,d_i)$はCC文d_iに出現する索引語t_jの出現頻度（出現回数）に基づいて，（2）式で算出する。

$$P(t_{js_j}\,|\,d_i)=\begin{cases}\dfrac{tf(t_j,d_i)}{\sum_{k=1}^{D}tf(t_j,d_k)} & if\ \ s_j=1\\[2ex] 1-\dfrac{tf(t_j,d_i)}{\sum_{k=1}^{D}tf(t_j,d_k)} & else\end{cases} \qquad (2)$$

ここで，$tf(t_j,d_i)$はCC文d_iにおける索引語t_jの出現頻度であり，DはCC文の総数である。

また，（1）式における$P(q\,|\,t_{1s_1},\cdots,t_{Ts_T})$は$t_{1s_1},\cdots,t_{Ts_T}$に固有名詞などが多く含まれるほど大きくなるように（3）式で定義する。つまり，さまざまな番組に出現する索引語と比較して，ある特定の番組だけに出現するような索引語はそのEPG文を特徴づけるうえで重要な役割を果たしていると考える。

$$P(q\,|\,t_{1s_1},\cdots,t_{Ts_T}) = \sum_{j=1}^{T} \frac{tf(t_j,q)\cdot r(t_{js_j})}{\sum_{k=1}^{T} tf(t_k,q)\cdot r(t_{k1})} \quad \text{——（3）}$$

ここで，$r(t_{js_j})$は命題t_{js_j}が成り立つことの希少価値を表し，数値が大きいほど希少価値が高い。提案手法では，索引語の希少価値を過去の放送番組のCCテキストにおける語のエントロピー$H(t_j)$に基づいて（4）式で定義する。

$$r(t_{js_j}) = \begin{cases} \log \sum_{k=1}^{N_v} tf(t_j,\boldsymbol{G}_k) - H(t_j) & if \quad s_j = 1 \\ 0 & else \end{cases} \quad \text{——（4）}$$

ここで，$tf(t_j,\boldsymbol{G}_k)$は過去の放送番組$\boldsymbol{G}_k$のCCテキストに含まれる索引語t_jの総数，N_vは希少価値の算出に使用する過去の放送番組の総数である。また，（4）式では，出現の偏りが大きいほど$r(t_{js_j})$の値が大きくなるようにエントロピー$H(t_j)$の増減を反転し，更に，$r(t_{js_j})$が負にならないようにしている。$H(t_j)$は（5）式で与えられる。

$$H(t_j) = -\sum_{i=1}^{N_v} \frac{tf(t_j,\boldsymbol{G}_i)}{\sum_{k=1}^{N_v} tf(t_j,\boldsymbol{G}_k)} \log \frac{tf(t_j,\boldsymbol{G}_i)}{\sum_{k=1}^{N_v} tf(t_j,\boldsymbol{G}_k)} \quad \text{——（5）}$$

提案手法では，実際の放送番組における索引語の出現頻度に基づいて類似度の算出に必要な係数$r(t_{js_j})$を決定するので，人手による不要語リストの作成や経験則に基づく重み付けなどの作業が不要で，より客観的な評価ができる。

3．実験

実際のテレビ番組を対象として，番組紹介映像の生成実験を行った。実験には自然ドキュメンタリー番組の「地球・ふしぎ大自然」を10本（回分）用いた。この番組は，毎回，ある動物や場所をテーマとした構成で，番組の音声はナレーションによる説明が主なものである。１本の番組の長さは43分で，１本あたり約350のCC文と約300のショットが含まれている。EPG文およびCC文の形態素解析には，形態素解析プログラム「茶筌（ちゃせん）」[13]を使用した。

3.1　番組紹介映像の生成結果

提案手法を用いて，EPGテキストの各文に対して最も類似したCC文を抽出した。また，抽出したCC文に対応する映像区間を決定し，その映像内容を調査した。**1表**に１例としてキイロスズメバチをテーマにした番組（EPG文は６文）に対する実験結果を示す。１表の映像内容の欄には対応する映像区間に含まれるショットのうちの最初のショットの内容を記述した。１表から，各EPG文に対して非常に類似したCC文が抽出できていることがわかる。番組紹介映像は番組の舞台となる信州の田園風景から始まり，軒下に作られた巨大な巣，巣を作るスズメバチのアップといった映像で構成されており，番組の内容が伝わる良好な結果となった。なお，生成された番組紹介映像の長さは37秒であった。

3.2　番組紹介映像の定量的な評価

提案手法の有効性を定量的に評価するために，専門家が制作した番組紹介映像を正解データとして映像内容を比較した。

3.2.1　評価の方法

番組紹介映像の内容は何をどのように紹介するかといった制作者の感性や主観によって異なる。従って，ある特定の番組紹介映像を絶対的な正解データとして評価する手法には問題がある。そこで，同一の番組に対して複数の番組紹介映像を作成し，それらをすべて正解データとした。複

1表　要約映像の生成結果

	EPG文	選択されたCC文	映像内容
1	民家の軒下などに巨大な巣を作るキイロスズメバチ。	今日は長野県の伊那谷で民家の軒下に巣を作るキイロスズメバチの謎に迫ります。	田園風景（８秒）
2	巣の直径は最大で１メートル、ハチの仲間では日本一の大きさだ。	ハチの仲間では日本一です。	軒下の巨大な巣（２秒）
3	巣作りのために立地の良い場所を探したり、建て増しや引っ越しまでしたり、人間の家造りそっくりの苦労がある。	巣作りのためには立地のよい場所を探したり建て増しや引っ越しまでしたり。	軒下を飛ぶハチ（７秒）
4	なぜそこまでして大きな家を作るのか？	なぜそこまでして大きな家を作るのでしょうか？	巣を作るハチ（４秒）
5	巨大な巣を作る背景には、家族の将来を見据えたある秘密が隠されていた。	キイロスズメバチがマンションのように大きな巣を作るのには家族の将来を見据えたある秘密があったのです。	軒下の巨大な巣（10秒）
6	信州・伊那谷でキイロスズメバチのある家族を徹底追跡し、すご腕建築術を明らかにする。	キイロスズメバチのすご腕建築術を驚くような映像で紹介します。	軒下の巨大な巣（６秒）

数の正解データと提案手法で作成した番組紹介映像の内容を比較することによって，より客観的な評価ができると考えた。また，正解データ間における映像内容の一致性を人手で調査することで，自動生成における番組紹介映像の精度を相対的に評価することもできると考えた。

実験では，２名の専門家が作成した30秒の番組紹介映像（正解１，正解２）と，放送で使用された30秒の番組スポット映像（正解３）の３種類を正解データとした。専門家２名にはEPGなどの情報は提示せず，番組本編だけを視聴して番組紹介映像を作成するように指示した。なお，放送で使用された番組スポット映像もほぼ同じ制作条件である。映像内容をショット単位で比較し，評価尺度には（6）式で表される擬似再現率*PR*（Pseudo−Recall ratio）と擬似適合率*PP*（Pseudo−Precision ratio）を用いた。

$$PR = \frac{N_r}{N_g},\quad PP = \frac{N_p}{N_o} \qquad \text{（6）}$$

ここで，N_gは正解に含まれるショットの総数，N_rは正解に含まれるショットのうち提案手法で抽出できたショットの数，N_oは提案手法で抽出したショットの総数，N_pは提案手法で抽出したショットのうち正解に含まれるショットの数である。通常の再現率や適合率＊5とは異なり，検出結果や正解映像においてショットの重複を許す指標である。

本実験におけるショットの一致判定について説明する。通常の番組制作では，１つの番組の中で，完全に同一の映像を複数の個所で使用することを演出上避けている。同様のシーンが必要となった場合には，カメラアングルや被写体の撮影サイズなどが似ている別の映像を使用する。そこで，実験では，ショットが画素単位で完全に一致していなくても，被写体の撮影サイズやカメラアングル，撮影内容などが演出的に同じであれば，一致していると判定した。

提案手法の有効性を評価するために，Taskiranらの手法[5]（３表の従来手法）と比較した。Taskiranらの手法は提案手法と同様に番組のナレーションに基づく要約手法であり，ドキュメンタリー番組に対して適用可能である。ただし，番組のEPG文などの情報は利用していない。Taskiranらの手法では，まず，不要語リストによって冠詞や形容詞，副詞などを取り除いた後，ナレーションにおける各文の重要度を算出する。次に，算出された重要度に基づいて，指定された要約映像の長さに収まるように文を選択し，最後に対応する映像区間を連結する。文の重要度は単語に基づく重要度と，文の散らばりに基づく重要度の重み付き和によって求めている。単語に基づく重要度は文に含まれる単語のTF−IDF（Term Frequency − Inverse Document Frequency）＊6と，番組内における特徴的な単語の共起がその文に幾つ含まれるかに基づいて算出する。また，散らばりに基づく重要度はエントロピーの増加量に基づいて，既に選択済みの文から離れた位置にある文が選択されるほど高くなるように算出する。なお，Taskiranらは音声認識を利用していたが，条件を公平にして比較するためにCC文を入力として用い，CC文を処理単位とした。また，生成する要約映像の長さを30秒に設定し，各種パラメーターはTaskiranらの論文に記載されている値を使用した。

更に，番組内容を考慮せずに機械的に映像を短縮する手法（３表の一定間隔）との比較を行った。これは，精度の下限を調査するためである。具体的には，10分間隔で６秒ずつの部分映像を取り出し，合計で30秒の映像を生成する。このサンプリング間隔は抽出される映像区間の数が提案手法とほぼ同程度となる値である。

3.2.2　実験結果

始めに，３種類の正解データの内容がどの程度一致しているかを調査した。**２表**に結果を示す。２表は制作者の違いによって，選択される映像区間がどの程度ばらついているのかを示している。*PR*と*PP*は共に38％〜53％であり，制作者が異なると，約50％は異なるショットが選択され，約50％は共通のショットが選択されるという結果であった。専門家によって制作された正解データ間の重なりが約50％であることから，提案手法においても約50％の精度を得ることができれば，ある程度，有効な手法であると判断できる。

提案手法および比較手法（従来手法，一定間隔）によって生成された番組紹介映像と３種類の正解データとを比較した。**３表**に結果を示す。提案手法の*PR*は平均で39％，*PP*は平均で43％であった。また，CC文だけを利用する従来手法の*PR*は平均で16％，*PP*は平均で24％であり，提案手法の方が高い精度であった。従来手法では，文に含まれる単語の数が多い場合や，番組内容とは関連の低い単語の共起などの影響で重要度が高く算出される場合に不必要なシーンが選択されることがあった。なお，機械的に映像区間を選択する手法（一定間隔）の*PR*と*PP*はいずれも平均で11％であり，最も低い精度であった。

3.3.　考察

提案手法による番組紹介映像は専門家によって制作され

＊５　計算式は同じであるが，検出結果や正解データに重複したショットを含まない指標。

＊６　単語の出現頻度と単語が含まれる文の数に基づいて単語の重要度を算出する手法。

2表　正解データ間の比較結果（10番組分）

	正解1		正解2		正解3		平均	
	PR	*PP*	*PR*	*PP*	*PR*	*PP*	*PR*	*PP*
正解1	−	−	38%	42%	44%	53%	41%	47%
正解2	42%	38%	−	−	41%	40%	41%	39%
正解3	53%	44%	41%	41%	−	−	47%	43%

3表　正解データとの比較結果（10番組分）

	正解1		正解2		正解3		平均	
	PR	*PP*	*PR*	*PP*	*PR*	*PP*	*PR*	*PP*
提案手法	39%	40%	36%	44%	40%	45%	39%	43%
従来手法	16%	25%	18%	25%	14%	23%	16%	24%
一定間隔	14%	13%	9%	11%	10%	9%	11%	11%

たEPGテキストに基づいており，映像の内容や順序にある程度のストーリーが感じられた。しかし，過去の放送番組など，EPGテキストが入手できない番組に対しては，提案手法は適用できない。このような番組に対しては，例えば，放送された番組スポット映像など，専門家が制作した番組紹介映像を正解データとして，番組紹介映像に適した映像・音声・言語の特徴を学習するような方法を検討する必要がある。

提案手法はドキュメンタリーや情報番組など，主としてナレーションで説明をする番組に対しては，比較的，良好な結果を得ることができる。しかし，ドラマやスポーツ中継番組など口語的な表現が多い番組では，EPGテキストとCC文の関連性やCC文と映像内容の関連性が低く，適切な映像区間を選択することが難しい。これらの番組に対しては，別のアプローチが必要である。

4．あとがき

電子番組表（EPG）に記載されている番組紹介テキスト（EPGテキスト）を利用して番組紹介映像を自動的に生成する手法を提案した。実際に放送された自然ドキュメンタリー番組に提案手法を適用し，専門家が作成した番組紹介映像と映像内容を比較した結果，擬似再現率は平均で39%，擬似適合率は平均で43%という結果が得られ，EPGを利用しないでクローズドキャプション（CC文）だけを利用する従来手法より良好な結果が得られた。

今後，過去の放送番組など，電子番組表が入手できない番組に対しても番組紹介映像を自動的に作成する手法の検討が必要である。更に，生成された番組紹介映像をより適切に評価する方法についても検討を進める予定である。

本稿は，電子情報通信学会論文誌に掲載された以下の論文を元に加筆，修正したものである。
河合，住吉，八木：“電子番組表における紹介文を利用した番組紹介映像の自動生成手法，”電子情報通信学会論文誌，Vol.J91−D, No.8，pp.2157−2165（2008），copyright

参考文献

1) N. Babaguchi, Y. Kawai and T. Kitahashi：“Generation of Personalized Abstract of Sports Video,” Proc. IEEE ICME ’01, pp.619–622（2001）

2) R. Lienhart, S. Pfeiffer and W. Effelsberg：“Video Abstracting,” Commun. ACM, Vol.40, No.12, pp.55–62（1997）

3) 森山，坂内：“ドラマ映像の心理的内容に基づいた要約映像の生成,”信学論（D–II），Vol.J84–D–II, No.6, pp.1122–1131（2001）

4) L. Agnihotri, K.V. Devera, T. McGee and N. Dimitrove：“Summarization of Video Programs Based on Closed Captions,” Proc. SPIE Conference on Storage and Retrieval for Media Databases, Vol.4315, pp.599–607（2001）

5) C.M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon and E.J. Delp：“Automated Video Program Summarization Using Speech Transcripts,” IEEE Trans. Multimedia, Vol.8, No.4, pp.775–791（2006）

6) Y. Rui, A. Gupta and A. Acero：“Automatically Extracting Highlights for TV Baseball Programs,” Proc. ACM Multimedia, pp.105–116（2000）

7) B. Li, H. Pan and I. Sezan：“A General Framework for Sports Video Summarization with Its Application to Soccer,” Proc. IEEE ICASSP ’03, pp.169–172（2003）

8) N. Babaguchi, Y. Kawai, T. Ogura and T. Kitahashi：“Personalized Abstraction of Broadcasted American Football Video by Highlight Selection,” IEEE Trans. Multimedia, Vol.6, No.4, pp.575–586（2004）

9) A.F. Smeaton, B. Lehane, N.E. O’Connor, C. Brady and G. Craig：“Automatically Selecting Shots for Action Movie Trailers,” Proc. ACM MIR ’06, pp.231–238（2006）

10) B. Ionescu, P. Lambert, D. Coquin, L. Ott and V. Buzuloiu：“Animation Movies Trailer Computation,” Proc. ACM, pp.631–634（2006）

11) W. Buntine：“A Guide to the Literature on Learning Probabilistic Networks from Data,” IEEE Trans. Knowledge and Data Engineering, Vol.8, No.2, pp.195–210（1996）

12) H. Turtle and W.B. Croft：“Evaluation of an Interface Network–Based Retrieval Model,” ACM Trans. Information Systems, Vol.9, No.3, pp.187–222（1991）

13) 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座（松本研究室）：“茶筌”，http://chasen–legacy.sourceforge.jp/

かわいよしひこ
河合吉彦

2001年入局。放送技術局を経て，2005年から放送技術研究所にてメディア処理の研究に従事。現在，放送技術研究所人間・情報科学研究部に所属。博士（工学）。

すみよしひでき
住吉英樹

1980年入局。広島放送局を経て，1984年から放送技術研究所にてコンピューターを応用した番組制作システム，メタデータ制作システムの研究に従事。現在，放送技術研究所人間・情報科学研究部専任研究員。博士（工学）。

やぎのぶゆき
八木伸行

1980年入局。甲府放送局，放送技術研究所，技術局，編成局を経て，現在，放送技術研究所研究企画部部長。画像・映像・メディア情報処理，コンピューターアーキテクチャー，コンテンツ制作技術，デジタル放送などの研究開発に従事。博士（工学）。

報告

蓄積されたニュース番組からの画像付きクイズ生成手法

佐野雅規　八木伸行　片山紀生†　佐藤真一†

A Method of Generating of Image-based Quizzes from News Video Archives

Masanori SANO, Nobuyuki YAGI, Norio KATAYAMA† and Shin'ichi SATOH†

要約

当所では，さまざまなコンテンツの開発や制作手法の研究開発を行っている。本稿では，コンテンツ制作技術の１つとして，ニュース番組からクイズを生成する手法を提案する。まず，クイズの多くのタイプから画像付き選択クイズを生成することに対象を絞り，問題の定式化を行う。具体的には，クイズに適した画像の選択，画像を説明している文の選択，似て非なる関係にある選択肢の生成の３つのサブタスクに分解し，各タスクに対する工学的アプローチを提案する。実験では，３つのタスクに対するそれぞれの評価と，それらを使って実現したクイズ生成という観点からの評価を行い，考察する。更に，実験結果を基に，今後，クイズ生成の精度を高めるためには何が必要な要素技術であるのかについて述べる。

ABSTRACT

We describe a method of generating quizzes from a news video archive. An image-based multiple-choice-quiz was formulated with three sub-tasks. These tasks include selecting an appropriate image in the quiz, selecting an appropriate sentence describing the image, and generating multiple choice questions on the image. We describe the engineering method for each sub-task and our tests of them. The effectiveness of the method was demonstrated in our experiments. Finally, we discuss what needs to be done to improve the accuracy and quality of generating quizzes.

†国立情報学研究所　National Institute of Informatics

1．まえがき

近年の目覚ましい情報処理技術の発展とインフラの急速な整備はコンテンツの制作にも大きな変化をもたらしている。NHKでは「3－Screens」というコンセプトを打ち出し，テレビ受信機向けだけではなく，携帯電話やパソコン向けのサービスを開始している。従来の番組制作のほかに，携帯電話やパソコン向けへのコンテンツの変換や，番組に関連するホームページ，デジタル放送におけるデータ放送用コンテンツ，ネットワークを利用した視聴者参加型のコンテンツの制作など，次々と新しいコンテンツが要求されるようになってきている。しかし一方で，コストを削減する必要もあり，コンテンツの制作作業を省力化するための自動あるいは半自動の制作支援システムが期待されている。

当所では，番組制作手法の研究開発の一環として，放送済みの番組を効率よく再利用するための研究を行っている。再利用するためには，まず，番組と番組に付随するさまざまなデータ（映像，音，テキストなど）を解析し，内容を記述する情報（メタデータ）を抽出する必要がある。また，メタデータを基にして，異なる型のコンテンツに変換する手法が必要である。メタデータを抽出するための研究では，多種多様なメタデータを効率よく付与するための仕組みであるメタデータ制作フレームワーク（MPF：Metadata Production Framwork）[1]を開発している。メタデータを活用するためのコンテンツ変換については，幾つかの番組を対象として研究開発を進めている。例えば，放送番組からのマルチメディア百科事典の自動生成[2]や番組紹介映像の自動生成[3]などがある。本稿では，番組を再利用するサービスとして，ニュース番組から画像付きクイズを生成するための手法を提案する。クイズというコンテンツは，昨今のゲーム機による学習，いわゆる，「脳トレ」の流行にみられるように，教育的コンテンツへの接触のモチベーションを高めることにもつながると考えている。また，インターネット上には，自分の得意分野についてのクイズを生成・公開して，みんなで楽しむことのできるサイト[4]も出現しており，クイズ生成は今後のコンテンツ制作の1つの大きな柱となると考えている。

本稿でのクイズ生成は全自動を目指すものではなく，半自動を目標としている。クイズは人間が知恵を絞って制作する創造物の1種であり，これを全自動で生成することは容易ではない。コンテンツの制作・公開を考えると，どのような形で制作が行われたとしても，最終的には人によるチェックが必要なので，半自動システムを目標とした。特に，放送局などにおいては，間違った情報を発信することは許されず，人によるコンテンツのチェックは必須である。このような背景から，始めに大量に蓄積された番組から計算機を使ってクイズ候補を自動生成し，その中から人手によって最終的に使用するものを選択することを想定する。本稿では，前段のクイズの自動生成を高品質で効率的に行う手法を提案する。

2．画像付きクイズの生成とは

2.1　素材と対象としたクイズ

ニュース番組を素材として，画像付きのクイズを生成する手法を提案する。

始めに，幾つかの画像付きクイズの例をあげ，必要な要素技術について述べる。1図は画像と文章を組み合わせたクイズのパターンを示している。(a) は画像に映っているものに関することを問う記述式の問題である。(b) も記述式であるが，前後の説明文をヒントに空欄を埋める方式にしている。(c) は説明文の真偽を問う問題，(d) は画像に最も適切な説明文を選ぶ選択問題である。これらのクイズを生成する手順を大きく分けると，画像の選択と画像に関連する文章の生成の2つになる。画像の選択におけるポイントはクイズに適した画像を選ぶことであり，このようなクイズの場合には，画像に映っているものが明確であるものが望ましい。特に，(a) や (d) は画像がなければ解けないクイズであり，文章との対応関係がより明確な画像でなければならない。文章の生成についてのポイントは以下のように整理できる。(a) の場合は，問題文そのものを生成する必要があり，これは幾つかのテンプレートを用意することである程度生成可能である。しかし，画像に何が映っているのかがわからないとテンプレートを選択できないので，何が映っているのかを特定することがポイントとなる。(a) 以外の文章については，字幕文（クローズドキャプション：視覚障害者向けにナレーションやせりふをテキスト化して放送する）を基に生成することができる。(b) の場合は，括弧抜きにする語彙（ごい）の選択がポイントである。ランダムに語彙を括弧抜きにするのは好ましくなく，字幕文中の重要な語彙を判定するなどの処理が必要である。(c) の場合も同様に，文中の重要な語を判定し，それを正解と間違いやすい語に置き換えることがポイントである。(d) の場合は，画像と文章の対応関係のほかに，複数ある選択肢が迷いやすい選択肢であることがポイントである。1図に示したクイズを生成するためには，クイズに適した画像の選択，被写体の特定，文章のテーマや重要部分の抽出，画像と文章の関係の抽出，解答に迷いやすい関係の抽出などが必要である。

本稿では，1図の (d) に示した選択クイズを対象とする。その理由は，そのクイズを生成する手法が他と比較し

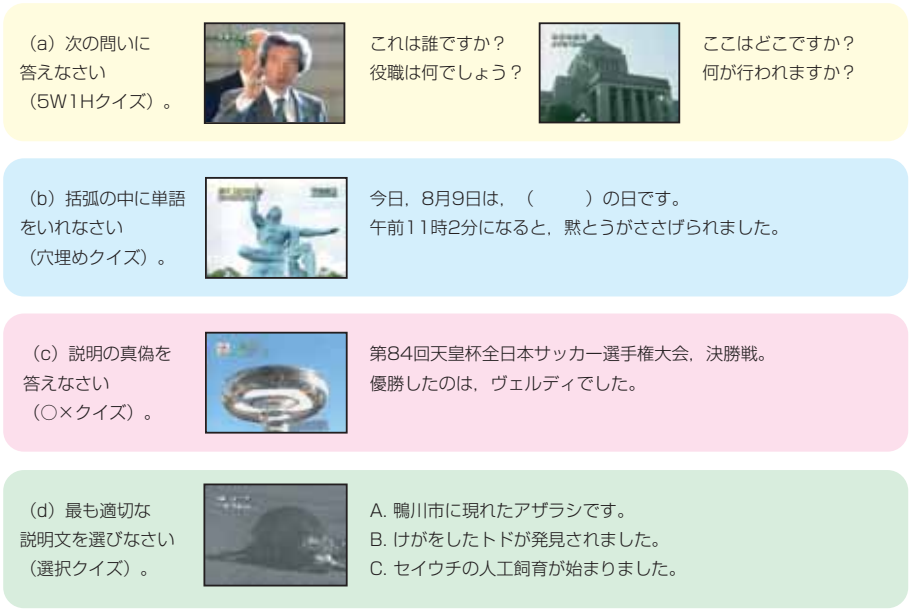
て簡略化でき，クイズ以外の役割，例えば，選択肢をすべて実際に起こったことで作成すればその素材となった番組へのナビゲーションという役割も付与することができ，蓄積された多くの番組に触れるきっかけを提供できると考えるからである。

次に，クイズ生成の難しさについて考える。クイズ生成というタスクを簡略化すると，ある事実が存在し，その一部分を隠し，そこを問う問題を作ることと言える。例えば，AとBの間にCという関係がある場合，「AとBの関係は？」あるいは「BとCという関係にあるものは？」という具合である。AとBは文章であっても，画像，音であってもよく，Cという関係もさまざまな定義ができる。従って，A，B，Cの組み合せによっては事実関係の抽出が非常に難しくなることがある。また，このような問題を生成しても，それがあまりにも常識的で簡単すぎる場合や，おもしろくない場合には，クイズになっていないと判断されることもあり，主観的な要素まで考慮に入れるとクイズを生成するということは難しい。更に，クイズの生成手法はクイズの役割・意図によって大きく異なる。例えば，単に知っているかどうか（知識の有無）を問うクイズもあれば，解くことで学習させることを目的としたクイズもある。また，おもしろさを埋め込み，笑いや納得を誘うエンターテイメント的なものもあり，同じ素材を用いたとしても，生成手法は大きく異なると考えられる。このようにクイズを生成するためには，さまざまな要素が絡み合う複雑で奥深い難しい問題があるが，本稿では，クイズを半自動で生成するための足がかりとして，比較的簡単な処理をした場合に何ができて何ができないのかということも明らかにする。

2.2 従来研究との比較

クイズや質問の生成を扱った論文が幾つかある。Higashinakaらは，ある人物に関連する文章を提示して，その人物を連想させるクイズについて報告している[5]。提示する文章がヒントであり，たくさんあるヒントをその連想の難しい順に並べ替える手法を提案している。並べ替えの対象である「人物に関連する文章」はあらかじめ与えられており，扱うデータもテキストだけである。また，ユーザーの質問に答えるだけでなく，システム側が自発的にユーザーに質問を行う情報提示システムの報告がある[6]。このシステムもテキストだけを扱い，例えば，あらかじめ作成しておいた京都に関する説明文の中の固有表現を隠す形で質問文を生成している。他の研究例として，ユーザーに情報を提示して理解させる場合には，辞書などを提示して単に読ませるのではなく，クイズ形式にして提示するのがより効果的であるという報告[7]や，職業別案内システムにおいて，お店の表現方法の違いを学習するためのデータを収集する手段として，クイズという枠組みを使い，ゲーム感覚で人々の参加を促してデータを収集するという報告[8]など，クイズというコンテンツの実社会への幅広い応用例が報告されている。本稿では，クイズの対象としてテキストと画像を扱い，素材から最終的なクイズ候補生成までの一貫した自動処理を提案する。

ここで，同じニュース素材を対象としたQA（質問・解答）システムとの比較を行う。QAシステムとクイズ生成



1図　画像付きクイズの例

システムは共にQとAを扱う点で似ている。しかし，QAシステムでは質問が与えられているのに対して，クイズ生成システムでは質問も生成する必要がある。また，QAシステムでは解答として正解だけを用意すればよいが，選択クイズ生成システムでは正解のほかに間違いやすい不正解も用意する必要がある。更に，QAシステムでは正解が明確であり，質問に対して必ず正解を解答しなければならないが，クイズ生成システムではこのクイズを絶対に生成しなければならないというようなことはない。例えば，ニュースの１項目に対して，必ず１つのクイズを生成しなければならないというようなこともなく，多くの素材から最終的に幾つかのクイズを生成することができ，それらがクイズになっていればよいということも異なる点である。このように，クイズを生成するというタスクは，これまでにない新しい研究要素を含んでいる。

3．提案手法

生成対象とした画像付き選択クイズは画像１枚と複数の文から成る選択肢で構成される。選択肢の１つの文は画像を説明した文であり，これが正解文である。それ以外の文を偽り文と呼び，正解文と似て非なる関係が求められる。このクイズの生成を以下の３つのサブタスクに分割して行う。

（1）クイズに適した画像の選択
（2）画像を説明した字幕文の選択
（3）似て非なる関係にある選択肢の生成

以下，各サブタスクをより具体的なタスクに置き換え，それをどのような手法で実行するのかについて述べる。

3.1　クイズに適した画像の選択

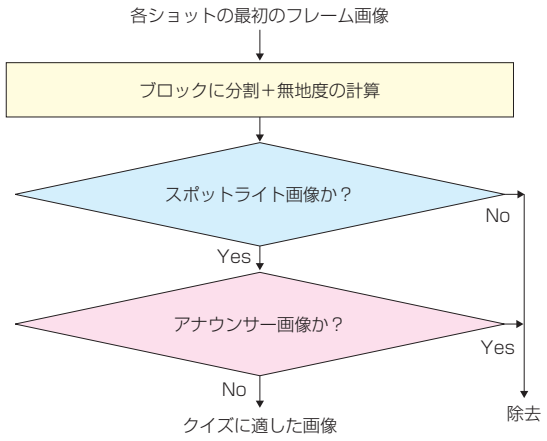
クイズに適した画像として，背景がほぼ無地で，中央に何かが映っている画像（以下，スポットライト画像と呼ぶ）を選択することとした。この理由は，今回の選択クイズでは画像と選択肢間の対応関係が重要で，解答者が画像の主となる被写体が何であるのかを容易に判断できるものが望ましいと考えたからである。なお，ニュース番組では，視聴者の意識を主被写体に向けるために，背景を無地にすることがある。

スポットライト画像を選択する手法を２図に示す。始めに，画像をサブブロックに分割し，各ブロックの無地の度合い（以下，無地度と呼ぶ）を数値化する*1。画像の左上のサブブロックの無地度の値から右下のサブブロックの無地度の値までを順にならべたものを無地度ベクトルと呼び，無地度ベクトルを特徴*2とした学習を行い，スポットライト画像かどうかを判定する。アナウンサーのバストショット（以下，アナバスと略す）はニュースの冒頭によくあるが，内容とは関係がないので，スポットライト画像から除去した。３図はスポットライト画像の例と無地度ベクトルの説明である。

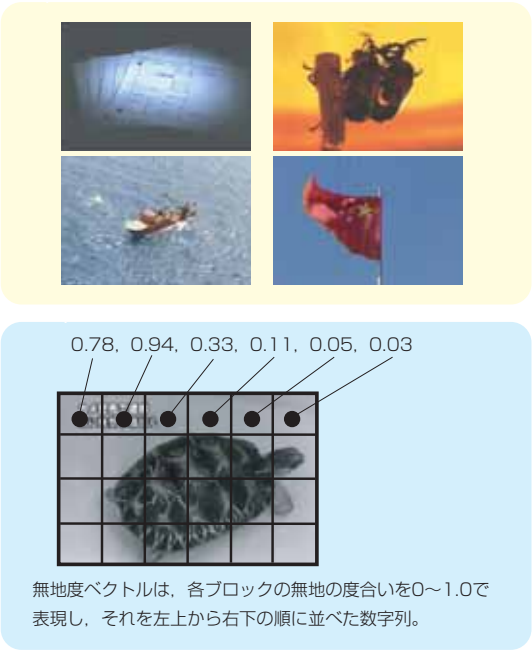
3.2　選択肢の生成

3.2.1　画像を説明した字幕文の選択

スポットライト画像についてのクイズの選択肢を生成する。正解文はその画像を含むニュース項目内の全字幕文の中から，画像と字幕文の特徴を用いて選択する。ここでの



２図　クイズに適した画像の選択手法



３図　スポットライト画像の例と無地度ベクトル

＊１　無地度の算出方法は4.1節参照。
＊２　スポットライト画像の無地度ベクトルには画面の周辺の数値が小さく（無地），画面の中央の値が大きい（何かがある）という分布特徴がある。

ニュース項目とは，一般的にトピックと呼ばれるニュース番組を構成する小区間で，通常，アナウンサーと共に表示されるトピックのタイトルで区切られる区間である。

4図に選択手順を示す。なお，字幕文を選択する前にスポットライト画像に対して顔検出を行う。もし，顔が検出された場合には，字幕文の選択の際に必ず人名が含まれることを条件に加えた。

始めに，正解文として画像が表示されている区間に発話されている字幕文を選択する。字幕文は発話されているときに表示されている画像と何らかの関係があるからである。画像が表示されている区間に字幕文がない場合には，画像の表示時刻の前後の字幕文で特徴的な言い回しの「このように」や「これが」という文節を含む字幕文を選択する。これらの言い回しは，アナウンサーが視聴者に映像を見るように促す場合に用いる言葉であり，字幕と画像の間により密接な関係があることを示唆しているからである。例えば，「仕組みは，このようになっています。」や「これが問題のゴミです。」というような文章である。特徴的な言い回しを含む字幕文が無い場合には，項目の冒頭部分から正解文を探す。項目の冒頭では，その項目の簡潔な説明（イントロ）を述べることが多く，冒頭部分の字幕文は項目内のすべての画像に関連があると考えるからである。以上の処理で，正解文を生成する。

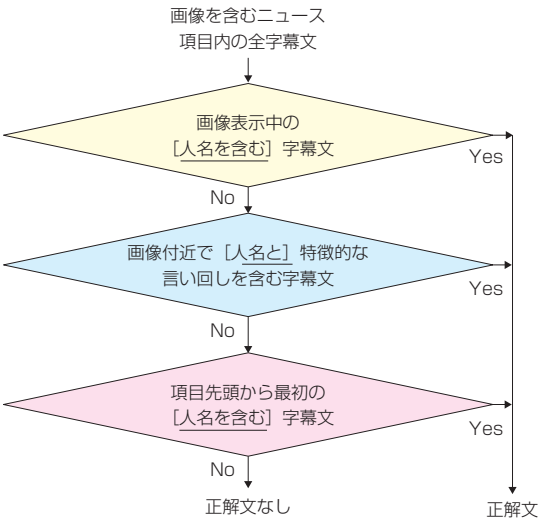
3.2.2　似て非なる関係にある選択肢の生成

似て非なる関係にある選択肢の生成手法を**5図**に示す。正解と選択肢間の似て非なる関係は各選択肢の文の主被写体と文全体の2つに注目して抽出する。

始めに，正解文から主被写体を特定する。正解文の中に人名が含まれていれば主被写体は人物とし，含まれていなければ文中の具象物名詞*3の中で，IDF（Inverse Document Frequency）値*4が最大の単語を主被写体とする。IDF値が高い単語とはある特定の文章だけに頻出する単語であり，一般的にその文章を特徴づけるとされている。なお，主被写体の抽出方法として「固有物名（Artifact）*5」を用いることも考えられるが，画面内の主被写体が必ずしも固有物名を持つとは限らないので，具象物名詞という枠で抽出する。

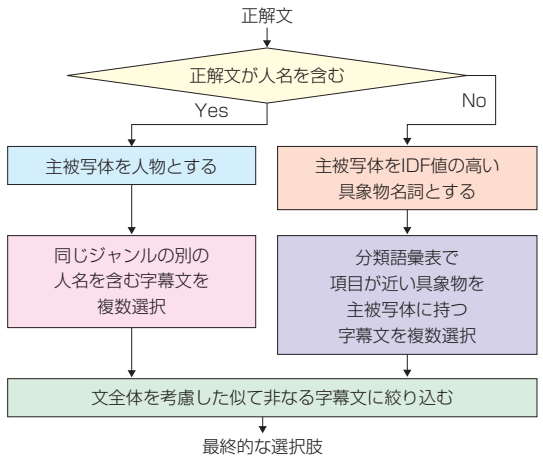
次に，主被写体に注目して似て非なる関係にある偽り文を選択する。偽り文は一定期間中のすべての項目内の字幕文を対象として，次の手順で選択する。主被写体が人物である場合には，同じジャンルの別の人物を主被写体とする字幕文を偽り文とする。例えば，同じスポーツ選手であるが違う人物や，同じ政治家であるが違う人物を主被写体とする字幕文を選択する。なお，ニュースの項目ごとにあらかじめスポーツ，国家関連（政治，国防，皇室など），その他の3つのジャンルに人手で分類した。主被写体が具象物の場合には，まず，当該期間中のすべての字幕文において，IDF値の高い具象物を主被写体として抽出し，その主被写体が類似している文を選択する。例えば，「イタリア」は分類語彙表の中で固有地名に分類されており，同じ分類項目内で違う単語の「スペイン」や「ポーランド」などが類似に該当する。以上の処理で，主被写体が人物である場合と無い場合について，それぞれ偽り文を複数抽出する。

最後に，複数抽出された偽り文の中から，更に，似て非



（画像中に顔が検出された場合は［　］の条件を追加）

4図　スポットライト画像を説明している字幕文の選択手法



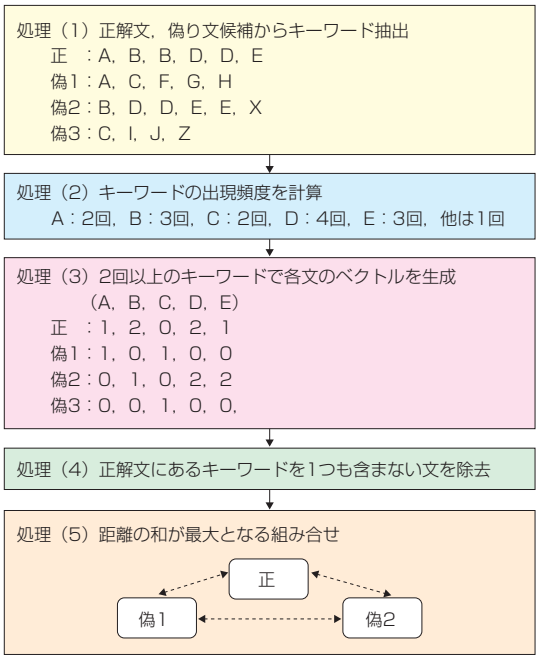
5図　似て非なる選択肢の生成手法

＊3　一般的に，目に見える形を持ったものを指す名詞。

＊4　単語が出てくるドキュメント数を総ドキュメント数で割った値。多くのドキュメントに出現する単語の重要度を下げ，偏って出現する珍しい単語の重要度を上げる指標。

＊5　IREX（Information Retrieval and Extraction Exercise：情報検索・情報抽出のための評価型ワークショップ）で定義された8つの固有表現（人名・地名・組織名・日付・時間・金額など）のうちの1つ。

なる関係の文を抽出し，最終的な選択肢を生成する。ここで，似て非なる関係の文とは「出現する単語は似ているが話題が違う関係」と定義した。ここでの話題とは，１つ１つの事件やイベントに相当する。従って，自動車事故のニュースでも，大阪で起こった事故と東京で起こった事故は違う話題として扱う。この処理を，出現する単語の共起パターン*6が違うと話題が異なるという仮説をたてて，６図に示す手順で実現した。まず，処理（１）では，正解文と偽り文のすべての文からキーワードとして名詞を抽出する。ただし，抽出する名詞の下位分類は，一般名詞や固有名詞などであり，代名詞や数，非自立，接尾，接続詞的な下位分類の名詞は抽出しない。処理（２）では，各名詞の出現頻度を調べ，処理（３）で２回以上出現したキーワードを要素とするベクトル（キーワードの出現した回数の組）を各文から生成する。処理（４）では，正解文に含まれるキーワードを１つも含まない文（図では偽３）を偽り文の候補から除外し，出現する単語が似ている偽り文だけを残す。処理（５）では，正解文と必要な数の偽り文すべての組み合せに対して，ベクトル距離の総和を求め，その値が最大の場合の組み合せを選択肢とする。距離が大きいということは，単語の共起パターンが違い，話題が違うということである。以上の処理で，似て非なる関係にある選択肢を生成する。



６図　似て非なる文の選択手法

＊６　複数の単語が同時に出現するパターン。例えば，単語AとBが同時に出現することは政治の話題で多いが，単語AとCが同時に出現することは経済の話題で多い，というようなパターン。

＊７　サポートベクターマシン（認識性能の高い学習モデルの１つで，教師あり学習を行う）を実装したプログラムの１つ。

＊８　サポートベクターマシンの中で，非線形問題を線形問題に置き換えるために用いる関数。

＊９　動径基底関数。任意の関数をその線形和で表すための関数で，SVMでは比較的よく用いられる。

＊10　離散コサイン変換の係数。

＊11　実際に検出できた正解データの数を全正解データの数で割った値。

＊12　検出したデータの中の正解データの割合。正解ではないデータ（ゴミデータ）が多く含まれるほど適合率は小さくなる。

４．実験結果と考察

4.1　クイズに適した画像選択の結果

NHKの「ニュース７」の2005年の１年分のニュースを用いて実験を行った。素材に用いた画像はすべてのニュース項目の各ショットの最初のフレームの画像で，全部で78,325枚である。最初のフレームの画像を用いたが，ニュース項目のショット内では構図の変化は小さく問題はない。以下，２図の手順に従って説明する。

始めに，各画像（320×208画素）を16×16画素のブロック260個に分割し，各ブロックで無地度を計算した。無地度の計算には，LIBSVM[9]*7を利用し，カーネル関数*8にはRBF（Radial Basis Function）*9を用いた。2004年12月のニュース項目の中からランダムに選択した45枚の画像をそれぞれ260個のブロックに分割し，無地ブロック6,080個と非無地ブロック5,620個に人手で分類したデータを用いて学習した。無地度はLIBSVMが出力する値，すなわち，評価対象のブロックが無地ブロックのクラスに属する確率を利用した。用いた特徴はブロックにおけるDCT係数*10のAC（交流）成分（255次元のベクトル）である。DC（直流）成分はブロックの平均輝度値であり，実際の無地の度合いはAC成分によって決まるからである。

スポットライト画像の判定にもLIBSVMを用いた。2004年12月のニュース項目の画像5,791枚をスポットライト画像249枚，非スポットライト画像5,542枚に人手で分類したデータを用いて学習した。用いた特徴は無地度ベクトルである。無地度ベクトルにはレイアウト情報（画像のどのあたりに，どの程度の無地があるかを示す情報）が含まれている。テストデータ78,325枚を用いて実験を行った結果，スポットライト画像として3,457枚が選択された。あらかじめ複数の人によってスポットライト画像かそれ以外の画像かを分類した画像でこの結果を評価すると，再現率*11は66.3％，適合率*12は63.8％であった。検出漏れや過剰検出された画像は検出されたスポットライト画像とほとんど同じ画像であった。判別器用の学習データおよび評価用の正解データを生成する際の主観評価にバラツキがあったことが原因ではないかと考えている。特に，背景の部分がほぼ無地の画像を正解とする場合に，「ほぼ無

地」と判断する基準は人によって大きく異なっていた。

アナバスの判定にも同じLIBSVMを用いた。スポットライト画像と判別された3,457枚の画像をアナバス（1,027枚）かそれ以外（2,430枚）の画像に人手で分類し，これを２分割の交差法*13で判断した。用いた特徴は各ブロック内のRGBの平均値を画像左上から右下まで順に並べたスタジオショットに特有の色の分布情報である。アナバスの判別結果は再現率が98.9%，適合率が99.5%であり，高い精度でアナバスの画像を除去することができた。すなわち，６枚のアナバスを除くことはできなかったが，最終的に，クイズに適した画像として2,436枚が得られた。

4.2　選択肢の生成結果

4.2.1　画像を説明した字幕文の選択

選択肢の生成実験はスポットライト画像として得られた2,436枚の中から，ランダムに250枚（約１/10）を選択して行った。４図に示すように，正解文の選択方法は画像中に顔が検出されたかどうかで処理が異なる。画像内の顔検出にはOpenCV[10] *14を用いた。実際に人物の有無に対する結果は再現率が81.5%，適合率が98.0%であり，手法としては良かったと考えている。次に，条件に合う字幕文を探す処理を行うが，人名などの判断には，南瓜[11]の固有表現抽出を利用した。１表は正解文の選択結果であり，画像250枚に対して212枚で正解文が選択できた。正解文が選択できなかったケースは画像に人が映っていると判断されたにもかかわらず，人名を含む字幕文がその項目の中に１つもなかった場合である。212の正解文が画像を説明した文になっているかどうかを検証した結果，99のケース（46.7%）で対応がとれていた。なお，この評価では５図で説明した主被写体が画像中で確認できるかどうかを判断基準とした。１表に示すように，人物については54.3%が対応していたが，人物以外の場合には27.9%であった。

4.2.2　似て非なる関係にある選択肢の生成

正解文が選択できた212枚の画像を対象として，抽出した主被写体（人物か具象物）に注目して，５図に示す方法で選択肢を生成（選択肢の数を３とした）した。結果を２表に示す。合計193のケースで選択肢が生成できていた。生成に失敗した理由は，人名も具象物も取り出せなかった場合と，６図の処理（４）で正解文に似ていない文を除去していった結果，偽り文がすべて無くなる，もしくは，１つになるケースであった。選択肢が生成できた193のケースの中で，似て非なる関係が認められたものは134（69.4%）であり，人物の場合には108（78.3%），具象物の場合には26（47.3%）であった。人物の場合で，似て非なる関係ができなかった原因の１つは，固有表現抽出で人名の判断を間違う場合でる。具象物では主に次の２つで失敗した。１つは複数の選択肢が同じ話題を含む場合で，今回の手法では話題を厳密に区別できていないことが原因である。他の１つは３つの選択肢の各話題にまったく共通する部分がなく，お互いに「似ている」という関係がない場合で，主被写体と特定した具象物が話題に直接関係のないことが原因であると考えている。

4.3　生成されたクイズの評価

３表は生成された193のクイズ候補について，画像付き選択クイズとなっているかどうかを人によって評価した結果である。

選択した正解文が画像と対応がとれており，生成した選択肢同士が似て非なる関係にあると判断できたクイズを筆者らが数えたところ，人物の場合で70，具象物の場合で12であった。これらについて，画像付き選択クイズになっているかどうかを評価すると，人物については70すべてがクイズになっていたが，具象物の場合には４であった。クイズになっているどうかを判断する基準は「画像と対応する選択文が１つだけあり，選択文の間に類似した関係がある場合」と定義した。人物の場合の１例をあげれば，画像中の人物の名前が選択肢１つだけに含まれており，他の選択肢には同職業の他の人名が含まれている場合などである。主被写体が具象物の場合も同じ基準で判断したが，３表に示すようにクイズになっていない場合が多かった。これは，映っている画像からは選択肢の似て非なる関係の対応が１つに絞れなかったことが原因である。例えば，選択肢はそれぞれノルウェーの国，ノルウェーの特定の人物，ノルウェーのレスキューチームについて記述しているが，画像にはノルウェーの国旗しか映っていないよう

1表　正解文選択の評価

	合計数	人物	人物以外
正解文選択成功	212	151	61
正解文が画像と対応	99 (46.7)	82 (54.3)	17 (27.9)

（　）内は割合（%）

2表　似て非なる選択肢生成の評価

	合計数	主被写体別	
		人物	具象物
選択肢の生成に成功	193	138	55
似て非なる関係成功	134 (69.4)	108 (78.3)	26 (47.3)

（　）内は割合（%）

*13　データを２つのセットに分割し，一方で学習を，他方でテストをする方法。学習とテストをするセットを逆にしても行い，その平均値を算出する。

*14　Intelが開発した画像処理用のC言語ライブラリー。

な場合である。3表の「最終的にクイズとして成功」は正解文の対応が不明確ではあったが，選択クイズとしては成功していると考えられるケース（7図のクイズ4で後述）を加えた結果である。3表は78（40.4%）の画像に対して自動でクイズが生成できていたことを示している。人物の場合には51.4%で成功しているが，具象物では成功率が12.7%であり，生成が難しいことを示している。

次に，クイズになっているかどうかの評価実験を被験者（放送コンテンツ制作の経験がない者）5人で行った。ただし，可能なかぎり通常の感覚で判断させるために，被験者には主被写体が何であるのかを明示せず，画像と文章との対応関係と，選択肢の間の似て非なる関係にだけ注目して判断をするように指示した。また，判断基準は同じで，画像と対応する選択文が1つだけであり，選択文の間に類似した関係がある場合をクイズになっていると判断した。3表の「5人中3人以上がクイズと判断」および「5人中4人以上がクイズと判断」した割合はそれぞれ50.8%と33.7%であった。また，いずれの場合もそのうちの約8割が人物のクイズであった。この結果は，筆者らによる判断と同じ傾向を示しており，ある程度の妥当性があると言える。被験者5人のκ（カッパ）値*15は0.45であり，評価結果はほぼ一致している。また，人物だけに限るとκ値は0.54，具象物だけに限ると0.12であり，具象物の場合の評

*15　主観評価結果の一致度を表す指標で－1～1の値を取る。1に近いほど一致度は高い。

3表　クイズとしての評価

合計数	主被写体別	主被写体別	
		人物	具象物
クイズとしての出力数	193	138	55
正解文の対応と似て非なる選択肢が成功	82	70	12
上段でクイズとして成功	74	70	4
最終的にクイズとして成功	78(40.4)	71(51.4)	7(12.7)
5人中3人以上がクイズと判断	98(50.8)	79(57.2)	19(34.5)
5人中4人以上がクイズと判断	65(33.7)	50(36.2)	5(9.1)

（　）内は割合（%）

クイズ1　（人物）

A. 女子シングルのフリーが行われ，前半2位の中野友加里選手が初出場で初優勝を決めました。
B. フィギュアスケート，グランプリシリーズの第2戦，スケートカナダは女子シングルのショートプログラムが行われ，村主章枝選手が2位でスタートしました。
C. 男子シングルの後半のフリーが行われ，18歳の織田信成選手が逆転で初優勝しました。

クイズ2　（人物）

A. また,内閣総理大臣，小泉純一郎と記帳をしていたのも今回は行いませんでした。
B. 佐々江局長は，今月17日にも予定されている，町村外務大臣と中国の李肇星外相との会談で,事態打開のための話し合いを行えるよう，中国側に強く促したいという立場を示しました。
C. 協議は，麻生総務大臣や竹中経済財政郵政民営化担当大臣ら関係閣僚が出席して午後5時から始まりました。

クイズ3　（具象物=素材）

A. そのうえで，従来のものより軽く，伸びたり縮んだりする素材を使い，通気性を良くして着心地が悪くならないよう，新たなくふうが加えられたということです。
B. また，今回の事故で車両がここまで大きく壊れたことについて電車の車両素材の強度や構造について詳しい立命館大学の坂根政男教授は次のように話しています。
C. 機体の耐熱タイルの傷が少なくとも26か所，また，操縦室の窓の近くの表面の傷やタイルの間から断熱素材が飛び出している個所，翼の前の部分の損傷などもありました。

クイズ4　（具象物=都会）

A. 実は今，都会の若者などに人気なんです。
B. 通りかかった買い物客などが，都会に出現した小さな尾瀬を楽しんでいました。
C. 打ち水で都会の暑さを和らげようという試みが東京浅草で行われました。

7図　生成されたクイズ例

価のばらつきが大きいことがわかった。

ここで，クイズとして成功した実例を示す。なお，ニュース番組ではニュース項目に関連する情報が画面に表示（スーパーインポーズ）されており，それが答えになっていることが多い。実例で示す画像にも情報が表示されているが，ここでは無視した。実際にクイズとして使用する際には，自動もしくは手動で情報を消すか，オリジナルの画像を使用することで解決できると考えている。

7図に主被写体が人物として生成されたクイズと具象物として生成されたクイズをそれぞれ2つずつ示した。なお，説明の都合で選択肢Aをすべて正解とした。

クイズ1は画像にはフィギュアスケート選手が，選択肢にはそれぞれ違ったフィギュアスケート選手の名前が含まれている。ただし，画像には女性が映っており，選択肢Cは男子の選手名なので，正解は選択肢AかBのどちらかという絞り込みができる。この情報はヒントであり，似て非なる関係の類似の程度を変えることでクイズの難易度を操作できることを示唆している。例えば，似て非なる関係の類似を「フィギュアスケート選手」ではなく，「女性のフィギュアスケート選手」にすることで，より難しいクイズを生成することができる。

クイズ2は画像には政治家が，選択肢にはそれぞれ違う政治家の名前が含まれている。画像には政治家の後ろ姿しか映っていないので，通常では誰なのか判断することは難しい。しかし，別の特徴（この場合は髪型）から，特定の人物と対応付けができるので，クイズになっている。これは，遊び心があるクイズ制作者が簡単には解けない問題を作成するために，何らかのヒントを利用して解くようなクイズを生成する場合に相当する。今回は偶然ではあるが，用いた表層的な解析技術のあいまいさ（この場合は画像の中央に何かが何となく映っていて，字幕文から人物と特定したこと）が，結果的に良かったと言える。

クイズ3は具象物のクイズとして成功した例である。具象物としては「素材」という単語が抽出されている。3つの選択肢にはそれぞれ，サッカーユニフォームの素材，電車の車両の素材，スペースシャトルの断熱素材が含まれており，素材という点で似ているが，それぞれ別物である。画像は素材のある部分を拡大しているので，3つの選択肢のどれにも当てはまるように思え，クイズになっている。ただし，画像の被写体は青色をしており，選択肢Aの「着心地」と「青色」で，勘の良い解答者であれば「サッカー日本代表のユニフォーム」との関係を見いだすことができるかもしれない。

クイズ4は正解文が明確に画像と対応しているとは言えないが，クイズになっていると判断したケースである。画像には，何かわからないものが映っており，選択肢を見ると，Aは「今，人気のもの」，Bは「尾瀬」，Cは「打ち水」を説明している。確かに，尾瀬でもなければ，打ち水でもなさそうであるが，クイズという見方をすると，画像のものが尾瀬の何かと，また，打ち水の何かと関係がないとは断言できない。もしかすると，尾瀬に出現した新種の何かか打ち水の効果を更に高める新素材かなどと解答者が考える可能性がある。実際はイルカの耳の骨であり，今，アクセサリーとして人気を呼んでいるというニュースの断片であった。クイズ4は興味をかきたてるクイズの例になっていると言える。数は少ないが，このようなケースが数件あった。

4.4　クイズ生成についての考察

クイズ生成の3つのサブタスクと，3つの組み合せによる生成手法について考察する。

始めに，クイズに適した画像の選択手法について考察する。今回は，スポットライト画像と表現した比較的無地の背景であって中央に何かが映っている画像に注目した。結果を見ると人物が多かったが，幾つものクイズが生成できており，主被写体のわかりやすいスポットライト画像がクイズに適した画像であるという仮定は間違っていなかったと言える。しかし，これ以外にもクイズに適した画像は存在する。例えば，中央に何かが映ってはいるが，背景が無地でないものもある。このような場合にも，画像中の主被写体の切り出し，言い換えるとセグメンテーションが正確にできれば，クイズ生成の精度は大きく上がる。また，ある特定の物体が主題ではなく，例えば，「祭り」のように，多くの人と特徴ある儀式などに象徴される画像もクイズになると考えられる。このような場合には，もはや画像だけで解析することは難しく，付随する字幕文など他の特徴を利用して判別する必要があると考えている。

次に，画像を説明している字幕文の選択手法について考察する。提案手法では，画像の被写体情報が人物であるかどうかを利用した。人物の場合には，字幕文を選択する際に人名に注目し，半数以上が正解であった。正解でなかったケースは画像中の人物がその項目の話題に直接関係なかった場合である。例えば，話題は国会での法案であるが，画像が政治家をシンボル的に映している場合や街角のインタビューを映している場合などである。従って，その話題に直接関係があるのかないのかという区別をすることも，今後の課題の1つである。一方，具象物の場合には，話題に関係なくIDF値などの統計的手法を用いて決めたので精度が低かった。画像に映っている被写体を幾つかのカテゴリーに分類するだけでも，クイズ生成の精度は上がると考えられ，今後，画像と言語処理の両面から解決したい

と考えている。

似て非なる関係にある選択肢の生成手法については，提案手法は人物の場合には有効であったが，具象物の場合には再検討が必要である。先にも述べたが，具象物の場合には画像と文章（主被写体）との対応がうまく抽出できなかったことが問題の根底と考えているが，仮に対応がうまく抽出できたとしても，以下のような問題のあることがわかった。それは，最適な「似て非なる」の関係を考慮できなかったことである。例えば，「さる」の似て非なる関係は類似点を「生物」と考えたときには，「人」「ライオン」「マグロ」「鶏」「細菌」など多種にわたるが，「ほ乳類」と考えれば，「人」や「ライオン」だけになる。どちらが良いのかは，そのクイズの目的に依存する。これについては，画像と言語のほかに，ニュース番組という対象データの特徴などを用いて解決すべき問題と考えている。

組み合せによるクイズ生成については，「画像」と「選択肢の似て非なる関係」の内容を考慮すべきだったと考えている。似て非なる関係の似ているものが画像に映っているとクイズにはならない。換言すれば，非なるものが画像に映っていなければならない。人物の場合には，被写体が特定できたので，この条件が明確に手順化できたと言える。画像の主被写体を特定するということは重要であり，ニュース番組の映像の修辞構造（文法）に注目することも１つの解決手法と考えている。

５．むすび

放送番組を再利用した新しいコンテンツの制作支援という目的で，ニュース番組から画像付き選択クイズを生成する手法について提案した。クイズ生成という課題をクイズに適した画像の選択，画像を説明した字幕文の選択，似て非なる関係にある選択肢の生成の３つのサブタスクに分解し，それぞれの手法を提案してクイズを生成した。その結果，250の選択画像から193のクイズ候補が自動生成でき，そのうちの78がクイズとして成り立っていることを確認した。78のクイズのうち，91％が人物に関するクイズであり，人物以外のクイズ生成の難しさが顕著に現れた。具象物では，画像中の被写体を特定すること（または，カテゴリーへ分類すること）が精度良くできなかったことが原因であり，今後，これらの技術を向上させる必要がある。また，被写体を特定する精度が上がれば，提案手法はニュース番組以外の，例えば，動物番組など他の番組にも適用可能と考えている。半自動でクイズコンテンツを制作することが目標であるとすれば，提案した手法は有効であり，今後，提案手法をベースとして，より内容の意味的な解析を考慮した研究を進める。また，将来的には，クイズの持つ「おもしろみ」や「ひねったクイズ」の生成手法などに挑戦したいと考えている。

本稿は電子情報通信学会論文誌に掲載された以下の論文を元に加筆・修正したものである。
佐野，八木，片山，佐藤：“蓄積されたニュース番組からの画像付きクイズ生成方法の提案，”電子情報通信学会論文誌，Vol. J92−D, No. １，pp.141−152（2009），

参考文献

1） メタデータ制作フレームワーク，http://www.nhk.or.jp/strl/mpf/
2） K.Miura, I.Yamada, H.Sumiyoshi and N.Yagi：“Automatic Generation of a Multimedia Encyclopedia from TV Programs by Using Closed Captions and Detecting Principal Video Objects,”IEEE Int. Symp. on Multimedia, pp.873–880（2006）
3） Y.Kawai, H.Sumiyoshi and N.Yagi：“Automated Production of TV Program Trailer Using Electronic Program Guide,”ACM Int. Conf. Image and Video Retrieval, pp.49–56（2007）
4） みんなの検定，http://minna.cert.yahoo.co.jp/
5） R.Higashinaka, K.Dohsaka and H.Isozaki：“Learning to Rank Definitions to Generate Quizzes for Interactive Information Presentation,”Annual Meeting of the Association for Computational Linguistics 2007, pp.117–120（2007）
6） T.Misu and T.Kawahara：“An Interactive Framework for Document Retrieval and Presentation with Question-answering Function in Restricted Domain,”Int. Conf. Industrial, Engineering & Other Applications of Applied Intelligent Systems, pp.126–134（2007）
7） R.Higashinaka, K.Dohasaka, S.Amano and H.Isozakia：“Effects of Quiz-style Information Presentation on User Understanding,”Interspeech 2007, pp.2725–2728（2007）
8） T.Paek, Y.-C. Ju and C.Meek：“People watcher：A Game for Eliciting Human-transcribed Data for Automated Directory Assistance,”Interspeech 2007, pp.1322–1325（2007）
9） C.-C Chang and C.-J. Lin：“LIBSVM：A Library for Support Vector Machines”（2001）Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/
10）OpenCV, http://opencv.willowgarage.com/wiki/
11）南瓜，http://www.chasen.org/~taku/software/cabocha/

佐野雅規（さのまさのり）

1994年入局。仙台放送局を経て，1997年から放送技術研究所にて，コンテンツ制作，メタデータ制作技術，メディア情報処理などの研究開発，ARIB,MPEG,EBUなどの標準化活動に従事。現在，放送技術研究所人間・情報科学研究部主任研究員。博士（情報学）。

八木伸行（やぎのぶゆき）

1980年入局。甲府放送局，放送技術研究所，技術局，編成局を経て，現在，放送技術研究所研究企画部部長。画像・映像・メディア情報処理，コンピューターアーキテクチャー，コンテンツ制作技術，デジタル放送などの研究開発に従事。博士（工学）。

片山紀生（かたやまのりお）

1995年学術情報センター助手。2000年国立情報学研究所助教授，2007年より准教授，現在に至る。データベースシステムに関する研究に従事。博士（工学）。

佐藤真一（さとうしんいち）

1992年学術情報センター助手。1995年～1997年カーネギーメロン大学客員研究員。1998年学術情報センター助教授，2000年国立情報学研究所助教授を経て，現在，国立情報学研究所教授。画像理解，画像データベース，映像データベースなどの研究に従事。博士（工学）。

情報還流システム

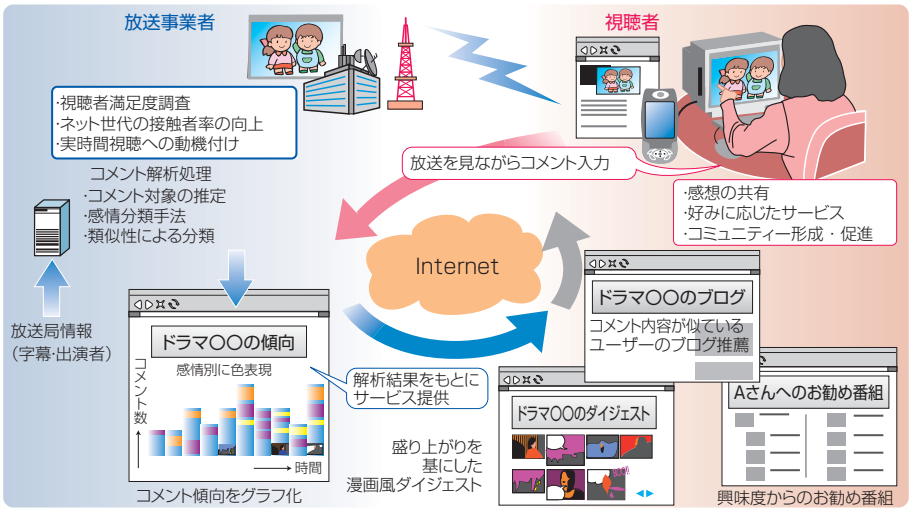
放送を通じた「巨大お茶の間」の形成

テレビはお茶の間の主役として家族団らんには欠かせないものであり，テレビ番組に関する話題が日常の会話のきっかけになることも多い。しかし，最近ではテレビを見ながらテレビ番組の感想やコメントをインターネットで送受信する人が若い世代を中心に増えている。そこで，当所では，放送局と視聴者や視聴者同士に新しいコミュニケーションの輪を広げることを目的として，インターネット空間の中に仮想的な「巨大お茶の間」を作ることのできる情報還流システムを提案している（1図）。

情報還流システムは，視聴者同士でコメントを共有するだけでなく，字幕放送や放送局にある出演者などの情報を使って，どのシーンのどの出演者に対するコメントかを解析し，視聴者の方にさまざまなサービスを提供することを特徴としている。提供するサービスの例として，集まったコメントの傾向をグラフ化し，その盛り上がりを基にした漫画風のダイジェストサービスや，コメント内容から抽出した興味度に合った番組お勧めサービス，コメント内容の似ている視聴者のブログを相互推薦するサービスなどを開発した。

これらのサービスを開発するための手法として，各コメントが番組のどのシーンのどの出演者に対するコメントかを推定する手法や，各コメントがどのような感情を表しているのかを肯定・否定・驚き・悲しみなどに分類する手法，コメント内容が似ている視聴者をグループ分けする手法，視聴者の好みを考慮したお勧めサービス生成手法などを提案している。

今後，１人でも多くの方にこれらの新しいサービスを楽しんでいただけるよう，「巨大お茶の間」に簡単に参加できる仕組みなどを開発し，放送に通信の機能を加えた新しい番組の楽しみ方を提案していく。



1図　情報還流システムの概要

スーパーハイビジョンの表色系

スーパーハイビジョンはより高い臨場感や質感の再現を目指したシステムである。色彩についても，実在する物体の色を忠実に再現するために，できるだけ広い範囲の色（広色域）を表現できることが望まれる。映像システムで色を表現する仕組みのことを「表色系」と呼ぶ。ここでは，スーパーハイビジョンの広色域表色系を紹介する。

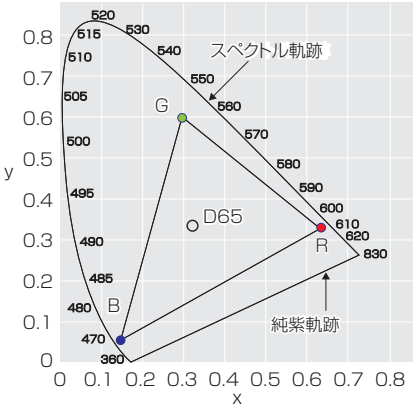
カラーテレビは赤（R），緑（G），青（B）の３原色の信号を足し合わせて色を再現している。従って，どのような色を３原色に選ぶかによって表現できる色の範囲が決まる。

ハイビジョンのRGB３原色の色度点を１図のxy色度図上に示す。馬てい形をした曲線はスペクトル軌跡と呼ばれ，レーザーのような単波長光源の色度に相当する。また，スペクトル軌跡の両端を結んだ直線を純紫軌跡と呼び，スペクトル軌跡に含まれない紫色を表している。人間が見ることのできる色はこの馬てい形の内側の色で，馬てい形の中心部から縁に向かって色の彩度が高くなる。ハイビジョンで再現できる色はRGB３原色の３点を結んだ三角形の内側の色だけである。また，ハイビジョンの基準白色は１図のD65*1と定められており，R,G,Bが同レベルのときに白色となる。

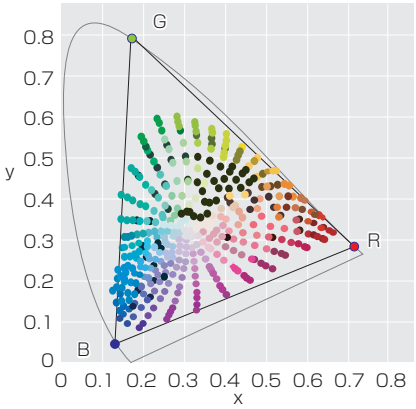
これまでのテレビは，ハイビジョンも含めて，主にCRTの蛍光体の特性の制約でRGB３原色が定められていた。しかし，近年，FPD（Flat Panel Display）技術が進歩し，CRTの色域を超えた色再現が可能になり，テレビ方式の広色域化の要求が高まっている。そこで，スーパーハイビジョンの色再現についてはCRTの制約にとらわれないかたちで検討を行った。その結果，他のメディアとの互換性，実在する色の再現性，コストと性能，デバイスやディスプレーの実現性の観点から，R，G，B共にスペクトル軌跡上の色を用いる新しい表色系を提案する。

２図にITU-R*2に提案しているスーパーハイビジョンの３原色色度点とポインターカラー*3と呼ばれる実在する表面色のデータベースの色度分布を示す。ハイビジョンでは，再現できない表面色が一部にあったが，提案の表色系ではほぼすべての色が再現できることがわかる。今後，スーパーハイビジョンの他の映像パラメーターについても検討し，臨場感の高い映像システムの実現を目指す。

＊１　CIE（国際照明委員会）が定める平均的な昼光色の標準の光。
＊２　国際電気通信連合無線通信部門。放送を含む無線通信技術の標準化機関。
＊３　実在する表面色の色域を表す測色データ。



1図　ハイビジョンの３原色



2図　スーパーハイビジョンの３原色とポインターカラー

特開2005332206　2005.12.02

映像イベント判別装置およびそのプログラム，ならびに，映像イベント判別用学習データ生成装置およびそのプログラム

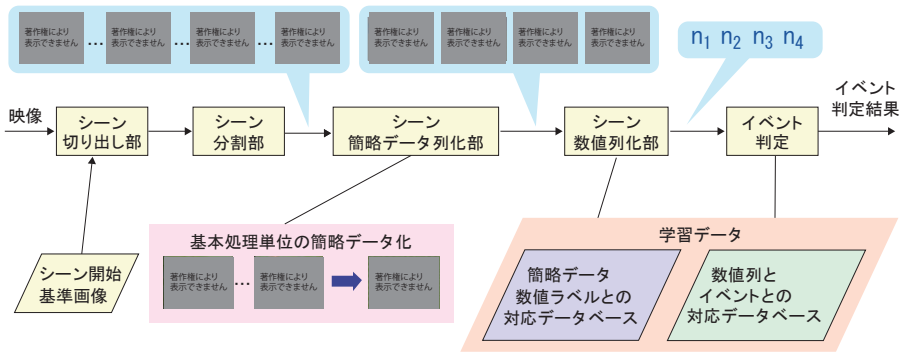
特徴と利用分野

本発明は，あらかじめ蓄積および学習しておいたシーンの映像特徴系列とそのシーン内で起こるイベントとの対応データを用いて，新しく入力された映像の各シーンで発生するイベントを自動的に判別する技術である。シーンの映像特徴系列を数値列に変換して処理するので，映像特徴量に付加される時間的あるいは空間的なノイズの影響を受けにくく，学習データの量と学習処理時間を削減することが可能である。カメラワークやスイッチングによる映像の推移とイベント内容との相関が強いスポーツ放送映像（野球や体操競技など）へイベント情報を自動的に付与することなどができる。

技術概要

本発明の構成を１図に示す。本装置は，映像ファイルを入力とし，シーン切り出し部，シーン分割部，シーン簡略データ列化部，シーン数値列化部およびイベント判定部で構成され，シーンのイベント判定結果を出力する。
シーン切り出し部では，シーンの開始に特有のフレーム画像で構成される「基準画像データ」を用いてシーン区間を切り出す。シーン分割部では，ショット切り換り点の自動検出技術などを用いて，シーンを基本処理単位へ分割する。シーン簡略データ列化部では，小領域のブロック追跡，画像特徴によるブロック統合処理などを行い，各基本処理単位を簡略データ化し，シーンを簡略データ列へ変換する。シーン数値列化部では，学習データ内の簡略データと数値ラベルとの対応データベースを用いて各簡略データを数値ラベルへ変換し，シーンを数値列で表す。最後のイベント判定部では，学習データ内の数値列とイベントとの対応データベースを用いて，数値列で表現されたシーンのイベントを判定し，出力する。

（発明者：望月貴裕，藤井真人）



1図　スポット映像生成装置の構成

特許第4456573号

映像抽出装置および映像抽出プログラム

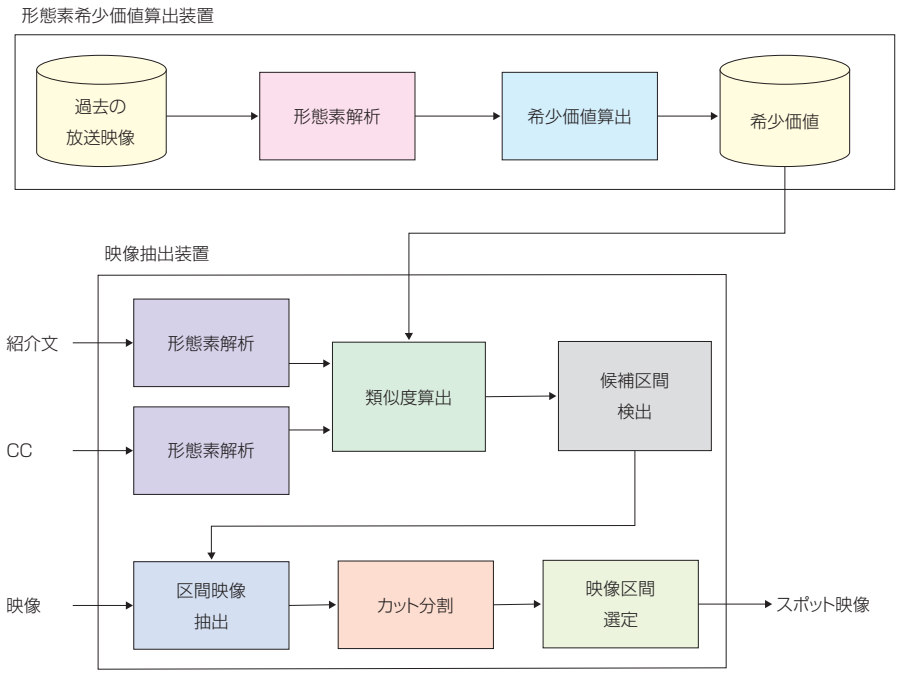
特徴と利用分野

本発明は，映像内容を説明するテキスト情報に基づいて，対応する映像区間を抽出する映像抽出装置に関する技術である。電子番組表などから入手した番組紹介文と，番組音声の書き起こしであるクローズドキャプション（CC）あるいは音声認識結果とを対応付けることで，番組紹介文に対応する映像区間を抽出する。本発明により，元の映像から番組内容を短く紹介するための番組スポット映像を自動生成することが可能となる。

技術概要

本発明の概要を1図に示す。本装置は，過去の放送番組の番組紹介文およびCCに含まれる形態素（テキストにおいて意味を持つ最小の単位）の希少価値を算出する形態素希少価値算出装置と，スポット映像を生成する映像抽出装置で構成される。形態素の希少価値は過去の放送番組における形態素の出現傾向（エントロピー）に基づいて算出し，番組紹介文とCC文の類似度を算出するための重み付けに利用する。映像抽出装置では，まず，番組紹介文とCC文の形態素を比較して番組紹介文に最も類似したCC文を選択し，選択されたCC文に対応する映像区間を抽出する。このとき，カメラの切り替え点やカメラ動きを考慮して，映像区間の時間範囲を調整する。最後に，抽出された映像区間を連結し，スポット映像として出力する。本装置を利用することで，元の映像を短く紹介するためのスポット映像を自動的に生成することが可能となる。

（発明者：河合吉彦，住吉英樹）



1図　本発明の構成

論文紹介

地上デジタル放送の放送波中継のための合成－比較－選択に基づく最ゆう判定指向型アダプティブアレー

電子情報通信学会論文誌，B, Vol.J91−B, No.9, pp.1072−1085（2008）

竹内知明，成清善一，横畑和典，今村浩一郎，濱住啓之，渋谷一彦

地上デジタル放送の放送波中継局における同一チャンネル干渉対策として，MMSE（Minimum Mean−Squared Error）アダプティブアレーを用いた干渉除去技術の適用を検討している。しかし，地上デジタル放送の放送方式であるISDB−T（Integrated Service Digital Broadcasting−Terrestrial）においてキャリヤー・シンボル空間に挿入されているスキャッタードパイロット（SP：Scattered Pilot）を参照信号とする従来の重み制御アルゴリズムには，SPの受信タイミングが希望波と一致もしくは近接しているISDB−T方式の干渉波を除去できないという問題があった。そこで，希望波と干渉波のSPの受信タイミング差にかかわらず干渉除去が可能な，ISDB−T用アダプティブアレーの合成－比較－選択に基づく最ゆう判定指向型重み制御アルゴリズムを提案する。計算機シミュレーションおよび野外実験を行った結果，提案法の有効性と地上デジタル放送の放送波中継局に最ゆう判定指向型アダプティブアレーを適用可能であることが確認できた。

赤色光増感型高感度15μm厚HARP光電変換膜の開発

映像情報メディア学会誌，Vol.62, No.12, pp.2031−2036（2008）

大川裕司，宮川和典※1，松原智樹，菊地健司，鈴木四郎，久保田節，谷岡健吉，小林　昭※2

※1 NHKエンジニアリングサービス　※2 浜松ホトニクス（株）

超高感度ハイビジョンカメラの実現を目指して，アモルファスセレン（a−Se）でのアバランシェ増倍現象を利用したHARP（High−gain Avalanche Rushing amorphous Photoconductor）光電変換膜（以後，HARP膜と呼ぶ）の研究を進めている。HARP膜はa−Seを主成分として構成されているが，a−Seのバンドギャップは約2.0eVであり，620nm以上の長波長の光に対してはほとんど感度がない。そのため，カラーカメラの赤色チャンネル用のHARP膜には，バンドギャップが0.34eVのテルル（Te）を増感材として添加している。Teの添加量を増やすと光電変換効率は向上するが，暗電流や残像が増加したり，膜欠陥が発生したりしやすくなるなどの問題があった。そこで，Teの添加量を増やしても特性劣化が生じないように，光入射側の界面付近に設けている電界緩和層の仕様を見直し，赤色光に対する光電変換効率を従来の２倍に高めた赤色光増感型HARP膜を開発した。

PVDF−driven flexible and transparent loudspeaker

Applied Acoustics, Vol. 70, No.8, pp.1021−1028（2009）

杉本岳大，小野一穂，安藤彰男，黒住幸一※1，原晃※2，森田雄一※2，三浦昭人※2

※1 NHKエンジニアリングサービス　※2 フォスター電機（株）

フレキシブルディスプレーや次世代のマルチチャンネル音響に適した薄型で柔軟なスピーカーの開発を進めている。今回，圧電性を持たせたポリフッ化ビニリデン（PVDF）で振動板のポリエーテルスルホン（PES）を駆動するシート状の透明なスピーカーを実現した。厚さ80μmのPVDFの両面に導電性高分子ポリチオフェンで透明電極を形成し，厚さ200μmのPESの背面全面にはり合わせて，Ａ４サイズの柔軟で透明なスピーカーを試作した。PVDFに電圧を印加して伸縮させ，振動板を駆動して発音させる仕組みである。一般に，圧電性のアクチュエーターにはヒステリシスに由来する無視できない歪(ひず)みがあるが，試作したスピーカーを１mの距離で音圧レベル約70dBSPLで再生し，PVDF単体の場合と比較して20~30dB低い高調波歪み率を実現した。この値は基本周波数応答レベルより50~60dB程度低い値であり，小型スピーカーの性能としては十分である。レーザードップラー振動計を用いて表面を観察した結果，試作したスピーカーの振動モードはPESに由来する振動モードが支配的であることがわかり，PVDFとPESを組み合わせたことで高調波歪みが低減したと推定した。

Multilevel Transitions of Closely Arranged Spin Valve Pillars Using Spin Transfer Switching

IEEE Transactions on Magnetics, Vol.44, No.11, pp.2519−2522（2008）

船橋信彦，町田賢司，青島賢一，宮本泰敬，河村紀一，久我　淳，清水直樹

超高精細・高速な空間光変調器の実現を目指し，スピン注入磁化反転を用いた光変調素子の研究を進めている。上部および下部電極を共有する２つの素子を近接配置したデバイスを作製してスピン注入磁化反転特性を測定した。素子サイズは0.3μm×0.1μmで，素子間距離を0.3μmから１μmまで数段階に変化させた。幅500msのパルス電流を印加した結果、いずれの素子間距離においても２段階の磁化反転を生じることがわかった。それぞれの磁化反転における抵抗変化分が等しいことから，中間状態では一方の素子だけが磁化反転していると推定した。また，反転電流密度の分布を調べた結果，素子間距離が１μmの場合に２つの素子は電極間に１素子のみを配置した素子と同じ反転特性を示したが，素子間距離を0.5μm以下に近接配置した場合にはより安定した中間状態をとることがわかった。２段階の磁化反転における安定した中間状態はスピン注入磁化反転を用いた超高密度MRAM（Magnetic Random Access Memory）の多値化や光変調素子における画素の多階調表示への応用が期待される。

学会発表論文一覧（2009年 7月～2009年 12月）

論文名	発表者	誌名	巻号
Optical Compensation of Distorted Data Image Caused by Interference Fringe Distortion in Holographic Data Storage	室井哲彦，木下延博，石井紀彦，上條晃司，清水直樹	Applied Optics	Vol.48, No.19, pp.3681–3690
字幕自動監視装置	椎名　努，成田長人，本間真一，今井　亨	映像情報メディア学会誌	Vol.63, No.7, pp.1006–1010
スーパーハイビジョンコーデックの開発と衛星伝送実験	井口和久，中島奈緒，境田慎一，合志清一（シャープ），筋誠　久，鈴木陽一，伊藤　隆（富士通研究所），中川　章（富士通研究所），数井君彦（富士通研究所）	電子情報通信学会論文誌，B	Vol.J92–B, No.7, pp.991–1002
ゴルフ中継での放送カメラを用いたティーショット軌道表示システム	高橋正樹，藤井真人，柴田正啓，八木伸行	電子情報通信学会論文誌，D	Vol.J92–D, No.7, pp.1036–1044
ARIBにおける高度衛星デジタル放送実証実験	橋本明記，井上康夫（WOWOW），松本英之（ソニー），方田勲（日立製作所），上田和也（パナソニック），市川鋼一（営電），佐藤　彰（NHKアイテック），柴田　豊（TBSテレビ），石原友和（テレビ朝日），太田陽介（フジテレビジョン），野崎秀人（テレビ東京），北之園展（スカパーJSAT），斉藤知弘，筋誠　久，小島政明，鈴木陽一，田中祥次	映像情報メディア学会誌	Vol.63, No.7, pp.957–966
Broadcast Trackback：テレビ放送コンテンツに対するユーザフィードバック機構	小俣拓也（慶應義塾大学），橋本範之（日本アイ・ビー・エム），重野　寛（慶應義塾大学），有安香子（次世代プラットフォーム），妹尾　宏	情報処理学会論文誌	Vol.50, No.7, pp.1735–1744
Fabrication of 5.8–inch OTFT–Driven Flexible Color AMOLED Display Using Dual Protection Scheme for Organic Semiconductor patterning	中嶋宜樹，武井達哉，都築俊満，鈴木充典，深川弘彦，山本敏裕，時任静士	Journal of the Society for Information Display	Vol.17, No.8, pp.629–634
Avalanche multiplication of photocarriers in nanometer–sized silicon dot layers	平野喜之，岡本健太（東京農工大学），山崎　晋（東京農工大学），越田信義（東京農工大学）	Applied Physics Letters	Vol.95, No.6, 063109
Heat treatment to suppress image defect occurrence in amorphous selenium avalanche multiplication photoconductive film with improved red–light sensitivity	大川裕司，宮川和典，松原智樹，菊地健司，鈴木四郎，谷岡健吉，久保田節，江上典文，小林　昭（浜松ホトニクス）	IEICE Electronics Express	Vol.6, No.15, pp.1118–1124
SAPHIRE（Scintillator Avalanche Photoconductor with High Resolution Emitter readout）for low dose x–ray imaging：Lag	Li, Dan（State University of New York），Zhao, Wei（State University of New York），難波正和，江上典文	Medical Physics	Vol.36, No.9, pp.4047–4058
通信ネットワークを利用した放送サービスにおける個人情報保護	中村晴幸，藤井亜里砂，大竹　剛，真島恵吾，藤田欣裕，今泉浩幸，谷本幸一（日立製作所），山田隆亮（日立製作所）	映像情報メディア学会誌	Vol.63, No.9, pp.1272–1285
Estimation of individualized head–related transfer function	松井健太郎，安藤彰男	Acoustical Science and Technology	Vol.30, No.5, pp.338–347
利用履歴を秘匿できるコンテンツ配信・課金方式	飛田孝幸（みずほ情報総研），山本博紀（中央大学），土井　洋（情報セキュリティ大学院大学），真島恵吾	情報処理学会論文誌	Vol.50, No.9, pp.2228–2242
Pre–enhancement for High Spatial Frequency in Holographic Memory	木下延博，室井哲彦，石井紀彦，上條晃司，清水直樹	Japanese Journal of Applied Physics	Vol.48, No.9 issue3, pp.09LA03.1–09LA03.4
動的3次元モデルを用いた映像制作手法と群衆シーンへの応用	久富健介，冨山仁博，片山美和，岩舘祐一，松永孝治，井藤良幸，石原　渉	SMPTE Motion Imaging Journal	Vol.118, No.7, pp.29–36
撮像系の空間周波数特性を維持した画像のグレア補正	正岡顕一郎，菅原正幸，野尻裕司，内川恵二（東京工業大学）	電子情報通信学会論文誌，A	Vol.J92–A, No.10, pp.669–676
Carrier–to–Noise Ratio in Magneto–Optic Transfer Readout Using Magnetic Garnet Film	野村龍男（静岡理工科大学），岸田雅彦，林　直人，岩崎勝男（FDK），梅澤浩光（FDK）	Journal of the Magnetics Society of Japan	Vol.33, No.6–2, pp.477–480
Application of Si–N Insulation Layer to CPP–GMR Device	船橋信彦，青島賢一，町田賢司，久我　淳，清水直樹，中川茂樹（東京工業大学）	Journal of the Magnetics Society of Japan	Vol.33, No.6–2, pp.521–524

学会発表論文一覧（2009年　7月～2009年　12月）

論文名	発表者	誌名	巻号
Stacked Image Sensor with Green– and Red –Sensitive Organic Photoconductive Films Applying Zinc–Oxide Thin Film Transistors to a Signal Readout Circuit	相原　聡，瀬尾北斗，難波正和，渡部俊久，大竹　浩，久保田節，江上典文，平松孝浩（高知工科大学），松田時宜（高知工科大学），古田　守（高知工科大学），新田浩士（高知工科大学），平尾　孝（高知工科大学）	IEEE Transactions on Electron Devices	Vol.56, No.11, pp.2570–2576
$SrGa_2S_4$：Bi蛍光体のフォトルミネッセンス特性	堺　俊克，田中　克，岡本信治	照明学会誌	Vol.93, No.11, pp.798–801
Low Voltage and Hysteresys–free, N–type Organic Thin Film Transistor and Complementary Inverter with Bilayer Gate Insulator	藤崎好英，假田雅史（東京工業大学），熊木大介，時任静士，山下敬郎（東京工業大学）	Japanese Journal of Applied Physics	Vol.48, No.11, pp.11504.1–111504.5
スーパーハイビジョン映像を用いた実時間ハイビジョン電子ズーム装置の開発	船津良平，塚本　拓（アストロデザイン），今村崇之（NHK–ES），山下嘗行，三谷公二，野尻裕司	映像情報メディア学会誌	Vol.63, No.12, pp.1868–1876
5.8–inch phosphorescent color AM–OLED display fabricated by ink–jet printing on plastic substrate	鈴木充典，深川弘彦，中嶋宜樹，都築俊満，武井達哉，山本敏裕，時任静士	Journal of the Society for Information Display	Vol.17, No.12, pp.1037–1042
A Novel Composite Right/Left–Handed Rectangular Waveguide with Tilted Corrugations and Its Application to Millimeter –wave Frequency–Scanning Antenna	岩崎　徹，鴨田浩和，九鬼孝夫	IEICE Transactions on Communications	Vol.E92–B, No.12, pp.3843–3849

研究会・年次大会等発表一覧（2010年 1月～2010年 2月）

題目	発表者	発表先／誌名	資料番号	発表年月日
有機光電変換素子におけるシロール誘導体の添加効果	小林諒平（埼玉大学）， 福田武司（埼玉大学）， 鎌田憲彦（埼玉大学）， 相原　聡，瀬尾北斗， 幡野　健（埼玉大学）， 照沼大陽（埼玉大学）	電子情報通信学会技術研究報告　OME 有機エレクトロニクス	Vol.109，No.359， OME200974， pp.4144	2010.01.05
音楽再生における感動評価と音の臨場感	大出訓史，安藤彰男， 谷口高士（大阪学院大学）	日本音響学会研究会資料　聴覚	Vol.40，No.1，H 20101， pp.16	2010.01.11
SUBJECTIVE ASSESSMENT OF CODED VIDEO QUALITY OF AVC/H.264 4：2：2 CONTRIBUTION ENCODER	中島奈緒，井口和久，境田慎一	International Workshop on Advanced Image Technology （IWAIT 2010）	p.133	2010.01.11 ~12
Personalization of Broadcast Programs using Synchronized Internet Content	松村欣司， Evans，Michael J （BBC Research and Development）， 鹿喰善明， McParland，Andrew （BBC Research and Development）	IEEE International Conference on Consumer Electronics（ICCE 2010）	4.15	2010.01.11 ~13
Development of a Prototype DataBroadcast Receiver with a HighQuality Voice Synthesizer	世木寛之，田高礼子，清山信正， 都木　徹	IEEE International Conference on Consumer Electronics（ICCE 2010）	9.14	2010.01.11 ~13
10Gbps Forward Error Correction System for 120GHzband Wireless Transmisson	岡部　聡，池田哲臣， 杉之下文康，正源和義， 枚田明彦（NTT）， 矢板　信（NTT）， 久々津直哉（NTT）， 門　勇一（NTT）	The 5th annual IEEE Radio&Wireless Symposium（RWS 2010）	WE2C3， pp.472475	2010.01.10 ~14
Fully Collusion Resistant Traitor Tracing Scheme with Constant Size of Private Key	小川一人，花岡悟一郎 （産業技術総合研究所）， 藤井亜里砂，大竹　剛， 今井秀樹（産業総合技術研究所）	暗号と情報セキュリティシンポジウム予稿集CD ROM （SCIS 2010）	3A41	2010.01.19 ~22
Consideration of a transport mechanism on broadcasting from the viewpoint of emerging hybrid content delivery systems	青木秀一，青木勝典，真島恵吾， 浜田浩行	Workshop on MMT（MPEG Media Trasnport）［ISO/IEC JTC1/SC29/WG11 （MPEG）91st meeting］	session1	2010.01.20
立体ホログラフィーを目指したMO光変調器	青島　賢一，船橋信彦，町田賢司， 久我　淳， 石橋隆幸（長岡技術技科大学）， 清水直樹	映像情報メディア学会技術報告コンシューマ エレクトロニクス研究会，マルチメディア ストレージ研究会［電子情報通信学会磁気記録・ 情報ストレージ研究会共催］	－	2010.01.21
SoftXRayCharged Vertical Electrets and Its Application to Electrostatic Transducers	本泉真人（東京大学）， 上野　藍（東京大学）， 萩原　啓， 鈴木雄二（東京大学）， 田島利文， 笠木伸英（東京大学）	IEEE International Conference on Micro Electro Mechanical Systems（MEMS 2010）	pp.635638	2010.01.24 ~28
電波テレビカメラ	九鬼孝夫，鴨田浩和，津持　純， 岩崎　徹	映像情報メディア学会技術報告	Vol.34，No.3，BCT 201017， pp.6570	2010.01.28
低電圧有機TFT駆動による5.8イ　ンチフレキシブル有 機ELディスプレイ	中嶋宜樹，武井達哉，藤崎好英， 深川弘彦，鈴木充典，本村玄一， 佐藤弘人，山本敏裕，時任静士	映像情報メディア学会技術報告	Vol.34，No.2，IDY 20107， pp.2528	2010.01.28
音素情報を利用したBICに基づくオンラ イン話者識別	奥　貴裕，佐藤庄衛，小林彰夫， 本間真一，今井　亨	情報処理学会研究報告　SLP　音声言語情報処理	Vol. 2010SLP 080，No.9	2010.02.05
TV Program Retrieval based on Summary using ngrambased Similarity Weighed by Named Entity	後藤　淳，住吉英樹，宮崎　勝， 田中英輝，柴田正啓， 相澤彰子（国立情報学研究所）	Proceedings of the 14th ACM International Conference on Intelligent User Interfaces（IUI 2010）	pp.411412	2010.02.07 ~10
ロードレースコースにおける時空間トレリス符号の 伝送実験	中川孝之，光山和彦，鵜澤史貴， 神原浩平，池田哲臣	映像情報メディア学会技術報告	Vol.34，No.5，BCT 201026， pp.912	2010.02.10
ロードレースコースにおけるLDPC符号化MIMO OFDM伝送実験	光山和彦，神原浩平，鵜澤史貴， 中川孝之，池田哲臣	映像情報メディア学会技術報告	Vol.34，No.5，BCT 201027， pp.1316	2010.02.10
地上デジタル放送ISDBTの繰り返し復号型 8ブランチダイバーシティ受信特性	成清善一，横畑和典，岡野正寛， 高田政幸	映像情報メディア学会技術報告	Vol.34，No.5，BCT 201029， pp.2326	2010.02.10
120GHz帯を用いたスーパーハイビジョン 野外伝送実験	岡部　聡，池田哲臣， 杉之下文康，正源和義， 枚田明彦（NTT）， 矢板　信（NTT）， 久々津直哉（NTT）， 門　勇一（NTT）	映像情報メディア学会技術報告	Vol.34，No.5，BCT 201030， pp.2730	2010.02.10

研究会・年次大会等発表一覧（2010年 1月～2010年 2月）

題目	発表者	発表先／誌名	資料番号	発表年月日
非圧縮SHV信号のOTU3フレームによる光伝送実験	中戸川剛，中村円香，小山田公之，尾中 寛（富士通），並木 周（産業技術総合研究所），浅見 徹（東京大学）	電子情報通信学会技術研究報告 IA インターネットアーキテクチャ	Vol.109, No.421, IA 2009–92, pp.59–64	2010.02.12
空気流入制限による薄型光ディスクの面振れ抑制効果の改善	梶山岳士，小出大一，高野善道，徳丸春樹，小名木伸晃（リコー），阿萬康知（リコー）	映像情報メディア学会技術報告	Vol.34, No.8, CE 2010–13, MMS 2010–13, pp.5–8	2010.02.19
Innovation of Digital Radio in Japan – Mobile Multimedia Broadcasting in VHF band–	高田政幸	RadioAsia 2010	Session 1 : Ready for Digital Future? Radio we know is Ending !	2010.02.22 ~24
Performance Evaluation of LDPC–MMSE–SIC System with Time Interleaving in Mobile Line–of–sight Environment	光山和彦，神原浩平，中川孝之，池田哲臣，大槻知明（慶應義塾大学）	International ITG Workshop on Smart Antennas（WSA2010）	pp.80–87	2010.02.23 ~24

NHK技研R&D　NO.121（2010年5月）

2010年5月15日発行

編集・発行	日本放送協会　放送技術研究所
	© 2010　日本放送協会
	〒157−8510　東京都世田谷区砧 1−10−11
	電話　03−5494−1125
	ホームページ　http://www.nhk.or.jp/strl/
制作	株式会社オーム社
	〒101−8460　東京都千代田区神田錦町3−1
	電話　03−3233−0641
デザイン・印刷	株式会社東京研文社
	〒162−0802　東京都新宿区改代町45

NHK技研
R&D
2010年5月通巻第121号
平成22年5月15日発行　編集・発行＝日本放送協会 放送技術研究所

ISSN 0914-7535
Printed in Japan