社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－CVIM－153（70）
2006／3／17

メディア認識結果へのルール適用による メタデータの自動生成とダイジェスト配信サービスへの適用

桑野 秀豪† 山田 智一† 川添 雄彦†

†日本電信電話株式会社サイバーソリューション研究所 〒239-0847 神奈川県横須賀市光の丘 1-1

E-mail: †{kuwano.hidetaka, yamada.tomokazu, kawazoe.katsuhiko}@lab.ntt.co.jp

あらまし ダイジェスト映像配信サービス向けのメタデータの自動生成システムとモバイル端末向けのメタデータアプリケーションを紹介する。開発したメタデータ生成システムは映像・音声などの複数のメディア認識結果にルールを適用することで、番組中の重要なシーンを自動抽出し、抽出したシーンを予め決められた時間長のダイジェスト映像として自動編集する。また、モバイル端末上にパラパラ漫画モードなど番組ダイジェストを速覧できるアプリケーションソフトを実装した。開発システムを用いて、野球、サッカーなどのライブスポーツ番組に対し、ダイジェスト映像を自動編集した後、モバイル端末への即時配信を実現し、システムの有効性を確認した。

キーワード メタデータ生成、映像認識、音声認識、自然言語処理、サーバ型放送

Automatic metadata generation by applying heuristic rules to the results of media analysis and its use to digest video distribution service

Hidetaka KUWANO† Tomokazu YAMADA† and Katsuhiko KAWAZOE†

†NTT Cyber Solutions Laboratories, NTT Corporation, 1-1 Hikarinooka, Yokosuka-shi, Kanagawa, 239-0847 Japan

E-mail: †{kuwano.hidetaka, yamada.tomokazu, kawazoe.katsuhiko}@lab.ntt.co.jp

Abstract This paper proposes a automatic metadata generation system for digest video distribution service. The system extracts significant scenes by applying heuristic rules to the results of media analyses. The system can also automatically produce digest videos of the desired duration. We also developed the novel application of digest viewing for mobile TV service. The application provides multiple viewing modes such as text mode, video comic mode, and digest video mode. We implemented the application on a mobile phone and can demonstrate the sports digest distribution service.

Keyword metadata generation, video analysis, audio analysis, natural language processing, server type broadcasting

1. はじめに

近年、サーバ型放送サービス等、新たな放送サービス市場の創出を目的とするブロードバンドネットワーク上での放送コンテンツとメタデータを配信する技術の検討が進んでいる[1,2]。メタデータを利用した映像視聴方法の一つに番組のダイジェスト視聴がある。好きな時に、好きな番組の内容を短時間に把握でき、時間にあまり余裕のない現代社会の生活スタイルにマッチした視聴方法である。家庭内のテレビでの視聴時は勿論、特に、家庭内の環境に比べ、時間に制約のある外出先等のモバイル環境においては、よりユーザニーズにマッチした番組視聴方法と考えられる。

ダイジェスト視聴を実現するためには、番組映像中のどの時間にどのようなシーンが含まれているかといった情報がメタデータとして記述されている必要がある。例えば、ニュース番組の場合、個々のニューストピックの順番や内容に関する情報、あるいは，サッカー、野球といったスポーツ中継番組であれば、シュートのシーン、ホームランのシーンが番組中のどのあたりの時間で起こったかといった情報になる。このようなメタデータの生成作業を全て人手で行うと膨大な手間が必要となる。メタデータを利用したコンテンツ流通サービスを費用対効果のあるものとして隆盛させていくためには、メタデータの生成作業を効率化、低コスト化することが重要な課題となる。

このような課題に対し、近年、映像・音声認識、言語処理といったメディア認識技術を利用し、メタデータを低コストに生成するための様々な検討が実施されている[3,4,5,6]。我々も、これまでに同様のアプローチで、制作済みの番組映像やライブ放送される番組を対象としたメタデータ生成の作業フローの提案、及び開発システムを用いた作業時間等の評価を行ってきた[7,8]。従来の検討では、ニュース向け、サッカー向け、野球向け等、ある程度番組内容に依存したメディア認識技術やシステム構成を検討・評価してきた。

本稿では、従来よりも番組内容や制作済み番組かライブ番組かといった放送形態の違いに対する汎用性を向上させたメディア認識方式やシステム実装方法を提案する。メデ

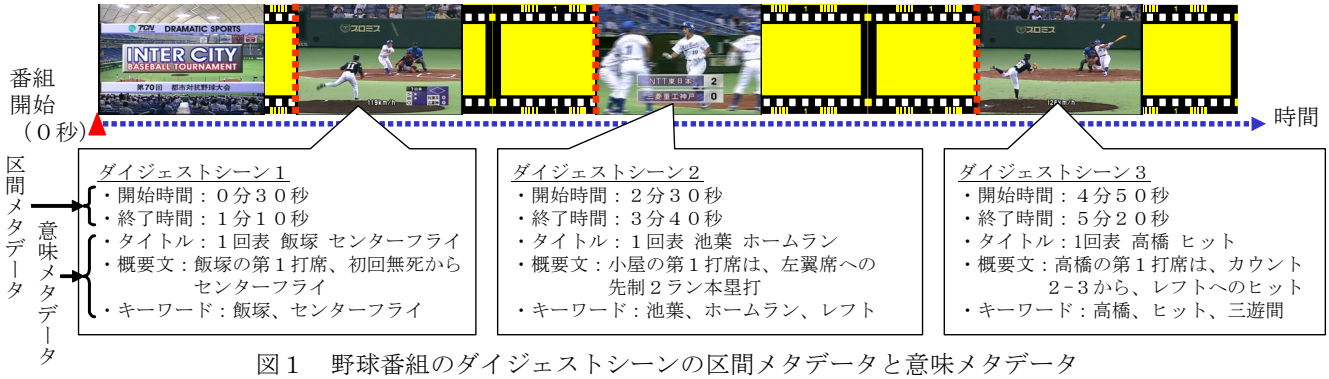

図１　野球番組のダイジェストシーンの区間メタデータと意味メタデータ

ィア認識方式に対しては複数種類のメディア認識結果の組み合わせに対しカスタマイズが容易なルールを適用することで、番組内容によらずダイジェスト映像に必要なメタデータを自動生成する方式を検討した。また、提案方式のサービス応用例として、モバイル端末向けのダイジェスト映像速覧インタフェースを紹介する。映像をパラパラ漫画のように閲覧し、短時間で映像内容のポイントが把握できるものである。

以降、2 章では本稿で扱うメタデータの定義を説明し、3 章でメディア認識結果へのルール適用によるメタデータ自動生成方式、4 章でメタデータ生成システムの実装方式、5 章で提案方式の実験結果と考察、6 章でモバイル端末向けの映像視聴インタフェース方式を説明し、7 章でまとめを述べる。

2. メタデータの定義

本稿におけるメタデータという言葉の意味を定義する。映像コンテンツに対するメタデータの国際標準規格が MPEG7[9]や TV Anytime Forum[10]等で策定されている。これらの国際標準規格では、番組中のシーンを「セグメント」と呼び、セグメントを説明する情報を「セグメントメタデータ」と呼んでいる。セグメントメタデータの内容には、そのセグメントの番組中における開始時間、終了時間、及びセグメントのタイトル、概要文、キーワードといったテキスト情報がある。以降、本稿では、説明の便宜上、前者を「区間メタデータ」、後者を「意味メタデータ」と呼び、議論をすすめる。図 1 に野球番組の区間メタデータ，意味メタデータの具体例を示す。

3. メディア認識結果へのルール適用処理

我々は、以前の検討[7,8]までは、特に区間メタデータについては、個々のメディア認識結果を手がかりに手作業での調整が必要な作業モデルを策定してきた。今回提案する区間メタデータと意味メタデータの自動生成方式は、複数種類のメディア認識結果の組み合わせパターンに対して、予め決めたルールを適用することで、個々のメディア認識結果を単独で扱う場合に比べ、より意味的に重要であり、サービスに直接利用できるメタデータを自動生成することが可能である。このような試みは従来も[6,17]の検討で行われている。本提案では、従来検討よりも映像系、音声系、言語系と互いに補完し合う様々なメディア認識エンジンを用いる。これにより、複数のメディア認識結果を組み合わせる場合も、シンプルでカスタマイズが容易なルールが策定でき、より汎用的な利用を可能とするものである。図 2 にその概念図を示す。以降、ルール適用処理例として、野球、サッカー、ニュースの各番組中の重要シーンの区間メタデータ、意味メタデータを抽出するルールを述べる。

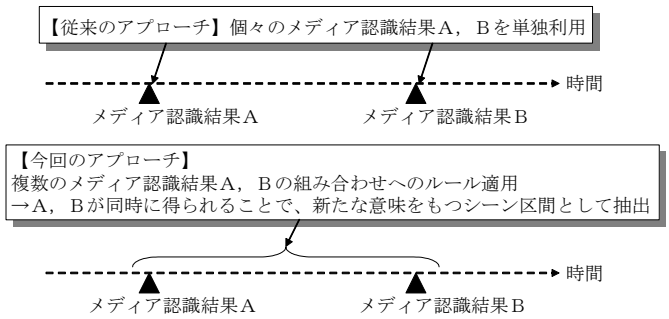

図２　メディア認識結果へのルール適用の概念図

3.1. 野球中継番組向けメタデータ生成ルール

野球番組のダイジェストには、得点シーンやファインプレーシーンが含まれることが望ましい。[8]の検討では、野球映像に対し、各バッターのシーンが終わったタイミングでオペレータがそのシーン内容に関するコメント文を発声する作業をベースとするメタデータ生成の作業モデルを考案、評価した。

今回は、図 3（a）に示すように、コメント音声入力時を重要シーンの終了時間とし、コメント音声入力時よりも前の時間で一番近い投球動作の検出時間を重要シーンの開始時間とするルールを策定した。投球動作は[11]の方法を用いて自動抽出可能である。[11]の方法は図 4 に示すような時間差分の履歴を反映した画像（Temporal Difference Image、以降 TDI）を利用するものであり、予め作成した参照用 TDI と入力映像から作り出す TDI とでマッチング処理を行う。また、図 5 に示すような得点時に表示される得点数字のテロップを検出することで、得点シーンを抽出することが可能となる。得点数字のテロップは番組中で毎回同じ大きさ、同じ位置に表示されることから、[12]の方法を用いて、選手名等の他のテロップと区別して検出できる。

また、図 3（b）に示すように、コメント音声、テロップ文字の認識結果を該当区間メタデータへの意味メタデータとして設定する言語処理ルールを策定した。これにより、一から手入力する必要なく意味メタデータの生成ができる。オペレータの発話コメント文は、「選手名、結果の順で話す」等特定の文法に則った発話ルールを設定することで、自由発話の実況音声に比べ、高精度な認識結果が得られる。

このように、動き検出、音声認識、テロップ文字認識、言語処理の各メディア認識処理結果の組み合わせに対し、野球向けのヒューリスティックなルールを適用することで、ダイジェスト制作に必要な重要シーンのメタデータの生成が可能となる。本ルールを実装する際は、利用するメディア認識処理を簡単に変更できるようなカスタマイズ可能な方式にすることで、野球以外の番組ジャンルにも適用可能としておくことがポイントである。

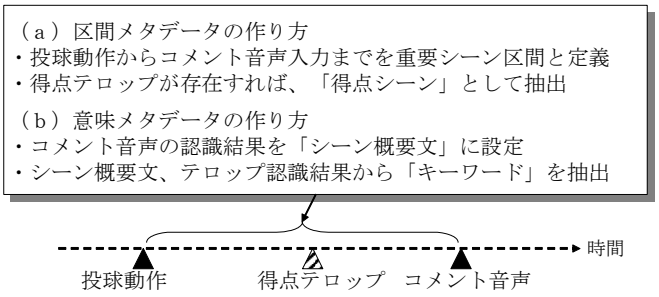

図３　野球中継番組向けのメタデータ生成ルール

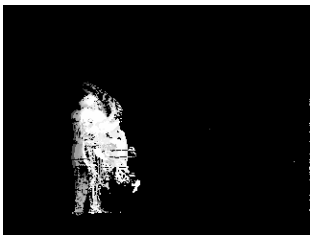

図４　ＴＤＩの例

図５　得点テロップの例

3.2. サッカー中継番組向けメタデータ生成ルール

野球の場合と同様にサッカー番組向けにゴール、シュートといったシーンのメタデータを生成するルールを策定した。図 6 に示すように、区間メタデータは、シュート前によく見られるゴール付近をズームインする等のカメラワークの検出結果[13]とシュート後に現れる選手名や得点数字のテロップ文字表示を利用したルールを策定した。また、シュート後に現れるリプレイ映像の開始時に使われる CG パターンを TDI を利用して検出することも可能である。意味メタデータは野球と同様のルールが適用可能である。

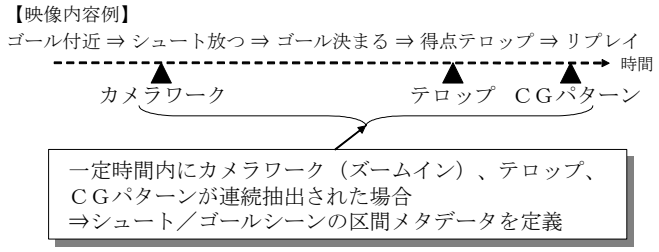

図６　サッカー中継番組向けのメタデータ生成ルール

3.3. ニュース番組向けメタデータ生成ルール

ニュース番組については、番組中の複数のニュストピックの区間メタデータ、意味メタデータを抽出するルールを策定した。図 7（a）に示すように、各ピックの冒頭に現れるニュースタイトルのテロップ文字を他のテロップと区別して抽出し、その直前のシーンチェンジの検出結果[13]を利用して区間メタデータを定義する。意味メタデータは、図 7（b）に示すように、[7]で紹介した方法を用いる。

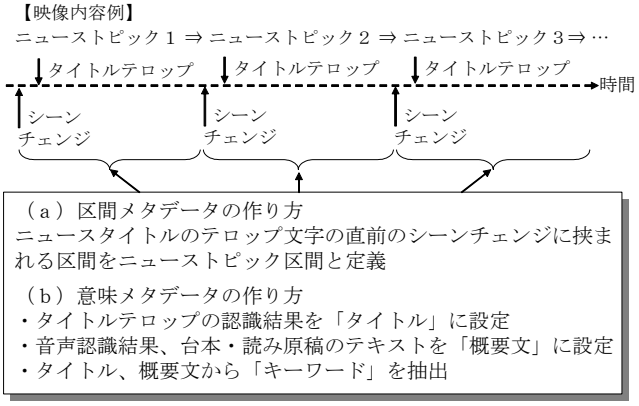

図７　ニュース番組向けのメタデータ生成ルール

3.4. 時間長指定ダイジェスト映像の自動編集方式

また、前節までに説明したようなルールを適用して抽出した番組中の重要シーンを集めて、予め指定した時間長のダイジェスト映像を自動編集する方法も検討した。図 8 に示すように、番組中から抽出した複数の重要シーンに対し、[14]の方法を適用し、音の盛り上がり度が高いシーンだけを選択し、1 本のダイジェスト映像として結合するものである。

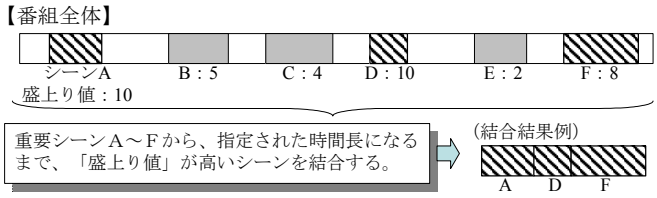

図８　ダイジェスト映像の自動編集

4. メタデータ自動生成システム

3 章で述べた方式を PC 上で動作するメタデータ自動生成システム『SceneCabinet/Live2』として実現した。本章では、その実装方法、及び実運用を考慮した各種機能を説明する。図 9 に SceneCabinet/Live2 のモジュール論理構成図を示す。3 章で述べた各ルールの実現に必要なメディア認識モジュール、ルール適用処理モジュールの他に、映像・メタデータ入出力モジュール、そして、メディア認識処理の結果を確認、編集するオペレータ操作用の GUI モジュールから構成される。

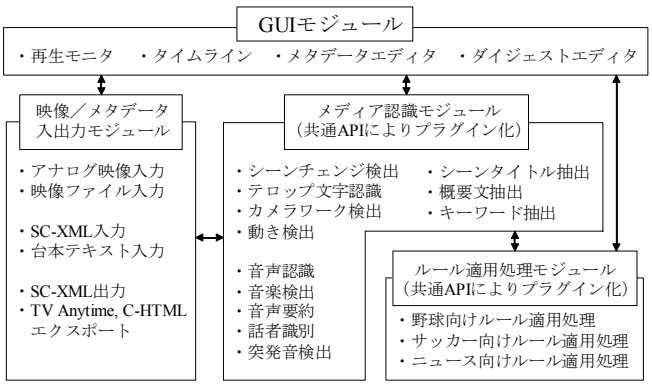

図 9　モジュール論理構成

4.1. 各モジュールの実装内容

メディア認識モジュールの中には、3 章で説明したメタデータ生成ルールで利用するシーンチェンジ検出、テロップ文字認識、動き検出、カメラワーク検出、音声認識、音声要約の他、音楽検出[15]、話者識別[16]、突発音検出[17]の映像・音声認識系の各モジュールが含まれる。また、シーンタイトル、概要、キーワードといった意味メタデータを自動抽出する言語処理系モジュールも含まれる。ルール適用処理モジュールには 3 章で述べた 3 種類の番組別のモジュールを実装した。メディア認識モジュールとルール適用処理モジュールはそれぞれで共通の API を規定し、プラグイン可能な DLL として実装した。これにより、将来的に新たなメディア認識やルール適用を追加する際も開発コストを抑えて実装することが可能となる。

また、映像入力モジュールには、アナログ入力とファイル入力の 2 つを実装した。アナログ入力モジュールは主にライブ番組の処理、ファイル入力は制作済みの番組映像の処理を想定して実装した。ファイル入力モジュールの対応フォーマットは MPEG1, MPEG2, WindowsMedia, QuickTime の各形式である。メタデータについては、各種メディア認識処理の中間結果等の詳細な情報も含めて記述できる独自の XML 形式（SC-XML 形式）を規定し、メタデータ編集工程時には、SC-XML 形式で保存、再読み込みなどを実施することを想定したものである。SC-XML 形式を各種映像配信サービス向けに利用する際の形式として TV Anytime 形式や特に携帯端末向けサービス用として C-HTML 形式に変換し、出力することもできる。

また、GUI モジュールとしては、再生モニタ、タイムライン、メタデータエディタ、ダイジェストエディタの 4 つから構成される。図 10 に SceneCabinet/Live2 のオペレーション画面例を示す。再生モニタはライブ放送向けに、アナログ映像をディジタイズしつつ、タイムシフト再生が可能である。タイムラインには、各種メディア認識モジュールやルール適用処理の結果の各種シーン情報が一覧表示される。メタデータエディタには各シーンの区間メタデータの調整機能、及び、意味メタデータの自動生成機能が付いている。ダイジェストエディタには、3.4 節で述べたダイジェスト自動生成機能が備わっている。

4.2. 各種モジュールのカスタマイズ

SceneCabinet/Live2 は、メディア認識モジュール、ルール適用処理モジュールや GUI モジュールを自由にカスタマイズできるよう設計した。例えば、メディア認識モジュールには 4.1 節の述べたように合計 10 種類以上の各種エンジンが備わっているが、常に全てのエンジンを同時実行すると、必要以上の情報が取得されたり、システム負荷も高くなる。例えば、3.1 節で述べたように野球番組の区間メタデータを生成する場合は、テロップ文字認識、動き検出、音声認識の結果が必須ではあるが、他のメディア認識処理は必須ではない。SceneCabinet/Live2 では、このように番組ジャンルに合わせて、各種モジュールの ON/OFF のパターンを定義することで、オーバースペックな処理実行を避け、最適なシステム負荷のもと処理実行ができる。

また、GUI モジュールにおいては、タイムラインに表示するシーン情報の表示名をカスタマイズできる。例えば、動き検出という一つのメディア認識の結果を野球番組の場合は、「投球動作」、サッカー映像の場合は、「リプレイ前 CG」といったそれぞれの作業フロー上、直感的に分かりやすい表現に簡単に変更できる。また、メタデータエディタにはメタデータ項目に合わせて、テキストボックス、ラジオボタン等の GUI 部品の選択、各部品の画面上の配置レイアウトをカスタマイズできる GUI ビルダー機能も備わっている。

以上のカスタマイズ機能により、SceneCabinet/Live2 は、メタデータ生成作業における対象番組ジャンルやメタデータ項目の変化に対して、システムの新規開発コストを抑え、汎用的な利用が可能となっている。

5. 実験結果と考察

3 章で述べたメディア認識結果へのルール適用によるメタデータ生成方式の精度評価を行った。野球、サッカー番組各 1 本、ニュース番組 3 本に対して、それぞれ 3.1 節、3.2 節、3.3 節で述べたルール適用処理での区間メタデータの生成精度を表 1 に示す。

表 1　ルール適用による区間メタデータの生成結果

	再現率	適合率
野球の各バッターシーン/得点シーン	87% （＝65 / 75）	100% （＝65 / 65）
サッカーのシュート/ゴールシーン	93% （＝25 / 27）	61%（＝25 / 41）
ニュース番組中の各ニュースストピック	100% （= 34 / 34）	100% （= 34 / 34）

表 1 より、ニュースに関しては、番組中の再現率、適合率とも 100%であり、そのままサービス利用できる高品質の区間メタデータが得られた。3.3 節のルールにおける、各ニュース冒頭のタイトルテロップや得点数字のテロップを全て正しく検出できたことが大きく寄与している。また、野球については、3.1 節で述べたルールにおける、TDI による投球開始タイミングの検出漏れが響き、再現率は 87%に

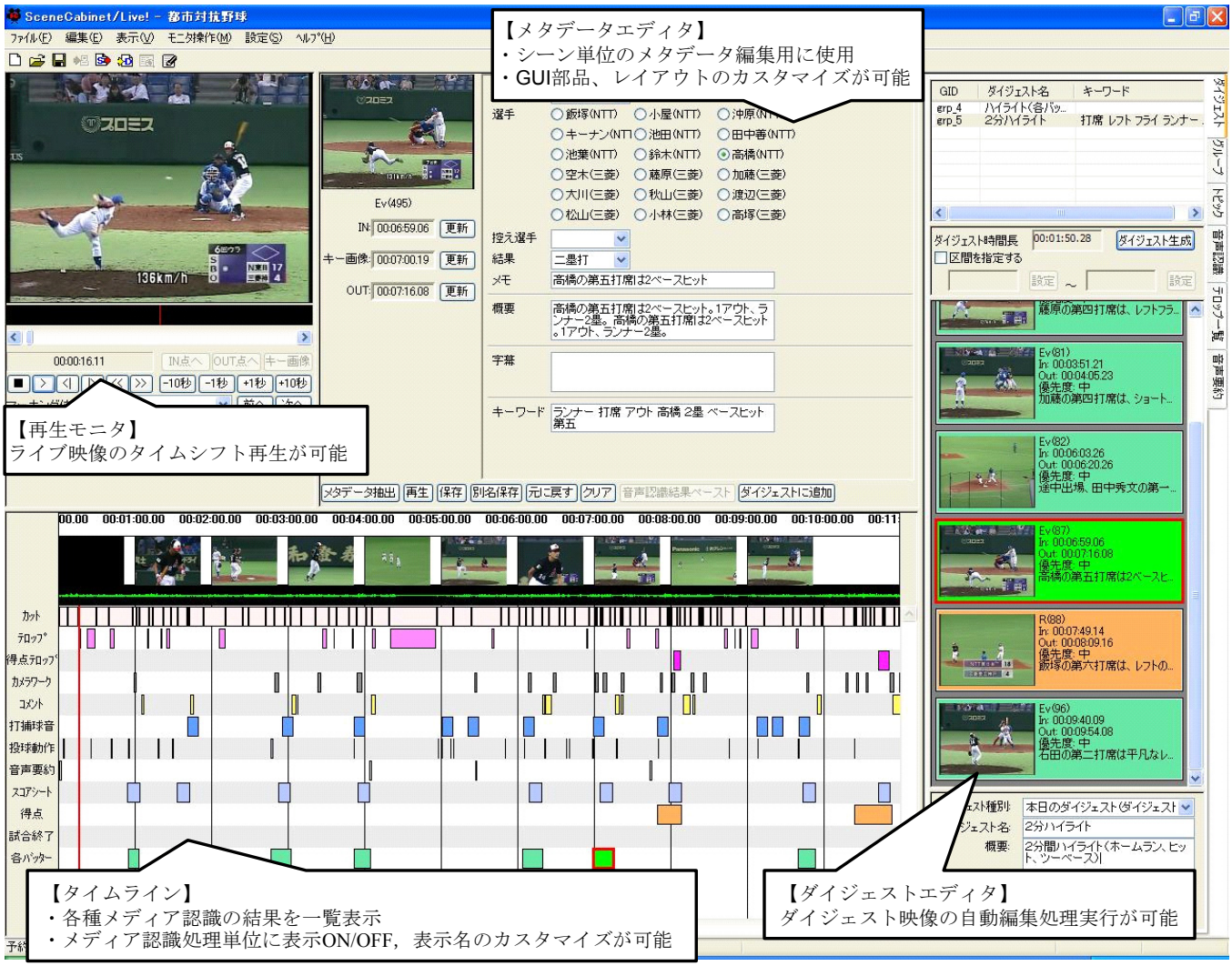

図１０　SceneCabinet/Live2 のオペレーション画面例

留まった。しかしながら、シーン終了時間としてのコメント発話入力時間から一定時間だけ遡った時間をシーン開始時間とすることで、投球開始タイミングが検出できなくても、簡単な手修正だけで区間メタデータの定義ができる。また、サッカーについては、リプレイ前 CG をほとんど漏れなく検出できたため、再現率は 93%と高かった。適合率が 61%とさほど高くないのは、ファールシーン等、シュート、ゴール以外のリプレイされるシーンを多数抽出したためであるが、サービス演出上必要な注目シーンとして、後工程で有効利用するケースも考えられる。

以上のように、今回検証した 3.1 節、3.2 節、3.3 節の各ルールでの区間メタデータの生成精度はいずれも一定以上のものであり、その効果が確認できた。ただし、いずれのルールもテロップ文字の検出を必要とするものであり、テロップが挿入される前の素材映像も含め、どんな映像に対しても汎用的に適用できるルールではない。そのような場合は、また別のルールを規定する必要がある。実運用時においては､映像にテロップが入っているかどうか以外にも、例えば、[7]で検討したような、映像内容に関する台本テキストが有効利用でき、メディア認識以外のアプローチが有効なケースなど、メタデータ生成の前提条件には様々なバリエーションがある。また､最終的なサービス仕様により、生成すべきメタデータの内容やその作業フローにも同様に様々なバリエーションがある。汎用性の高い方式を確立するのは勿論重要だが、それだけでなく、メタデータ生成の前提条件の変化に対して、毎回、新規検討・開発を行うのではなく、カスタマイズ容易性を考慮した方式、システム実装もメタデータ生成の低コスト化には重要と考える。今回の提案システムはその一例を示せたものと考える。

6. ダイジェスト配信サービスへの適用

SceneCabinet/Live2 で作成したダイジェスト映像のメタデータを利用したモバイル端末向けのダイジェスト視聴用アプリケーションを紹介する。モバイル環境では､「外出中で時間がないので映像内容のポイントだけ素早く把握したい」「音が聞き取りにくい」「画面が小さい」等、家庭内でのテレビ視聴時に比べ、映像視聴条件の制約が多い。このような制約下でも利便性の高い映像視聴ができるアプリケーションソフトをモバイル端末上に実装した。

本アプリケーションは、SceneCabinet/Live2 の出力するメタデータ、及び C-HTML を読み込んで、図 11 に示すように、目次モード、パラパラ漫画モード、映像モードの 3 つの視聴モードを提供する。目次モードは番組中のダイジェストシーンのタイトル、概要文といったテキストのみで映像全体を一覧できる。パラパラ漫画モードはシーンの静止画と説明テキストを 1 ページずつパラパラ捲りながら視聴する。映像モードは通常再生を行う。3 つのモードは連動しており、シーン単位でモードを切り替えて、ジャンプ再生する等の操作が行える。目次モードとパラパラ漫画モードは映像モードに比べ、短時間で映像全体のポイントだけを把握することができる。また、データ容量も映像データに比べ、非常に少なく、コンテンツ配信時の通信トラフィックの観点からも有効な方式である。

利用者はこれら 3 つのモードを、状況に合わせて選択して映像鑑賞ができる。また、外出中にお気に入りのシーンにブックマークを付けておき、それをメタデータとして保存し、帰宅後にメタデータをテレビに転送して続きのシーンから鑑賞するといったサービスも想定できる。このような固定・携帯連携サービスに必要な技術を今後開発していく。

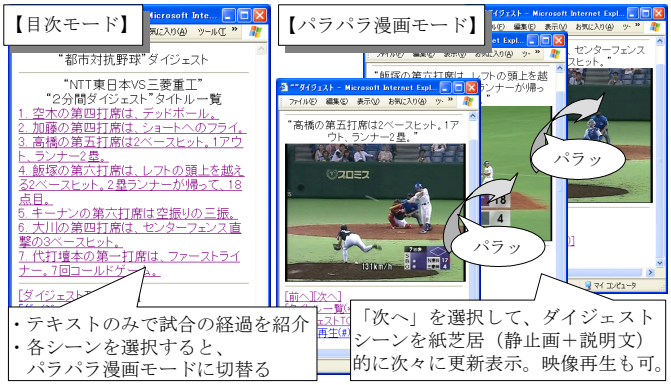

図１１　モバイル向け映像速覧アプリケーションの画面例

7. まとめ

複数のメディア認識結果にルールを適用することで、番組中の重要なシーンを自動編集し、ダイジェスト映像を制作するメタデータ生成システムとモバイル端末向けのメタデータアプリケーションを紹介した。実験においては、ルール適用によりニュース、野球、サッカーの各番組に対し、高品質なダイジェスト映像向けのメタデータが生成できることを確認した。今後は、メディア認識結果の学習によるルール策定も含めた一層の低コスト化、コンテンツ流通サービスシステムとの連携機能の開発を実施していく。

謝辞

本研究を進めるにあたり、日頃から有益な議論を頂いた NTT サイバーソリューション研究所の森本正志主幹研究員、長田秀信氏、三上弾氏、東正造氏、日高浩太氏、NTT サイバースペース研究所の藤井憲作氏に深く感謝致します。

文　献

[1] 電波産業会（ARIB）標準規格, “サーバ型放送における符号化，伝送及び蓄積制御方式,” ARIB STD-B38, 2004.

[2] 亀山渉, 花村剛, “デジタル放送教科書,” インプレス, 2003.

[3] 佐野雅規, 住吉英樹, 八木伸行, “サッカー中継における会場音とスピーチを利用したメタデータ生成,” 信学技報, PRMU2005-120, pp.33-38, Nov.2005.

[4] 三須俊彦, 高橋正樹, 藤井真人, 八木伸行, “スポーツ番組におけるメタデータ自動生成～サッカー選手追跡・同定のためのデータフュージョン～,” 信学技報, PRMU2005-121, pp.39-44, 2005.

[5] 木村雅之, 山内真樹, 大宮淳, 福宮英二, 西川順二, “知覚的クラスタリングに基づくスポーツ映像の自動インデクシング,” 信学技報, PRMU2005-121, pp.45-49, Nov.2005.

[6] 佐野雅規, 住吉英樹, 八木伸行, “情報統合機能を実装したメタデータエディタ,” FIT2005, I-028, pp.69-70, 2005.

[7] H. Kuwano, Y. Matsuo, and K. Kawazoe, “SceneCabinet: Semantic Metadata Extraction System combining Video/Audio Indexing and Natural Language Processing Techniques,” Proc. IBC2004, pp.458-466, Sept. 2004.

[8] H. Kuwano, Y. Kon’ya, T. Yamada and K. Kawazoe, “SceneCabinet/Live!: Real-Time Generation of Semantic Metadata combining Media Analysis and User Interface Technologies, ” Proc. IBC2005, pp.253-260, Sept.2005.

[9] MPGE7, http://www.chiariglione.org/mpeg/

[10] TV-Anytime Forum, http://www.tv-anytime.org

[11] K. Fujii and K. Arakawa, “Video Editing based on Motion Recognition using Temporal Templates,” Proc. ACM User Interface Software Technology 2003, pp.71-72, 2003.

[12] H. Kuwano, Y. Taniguchi, H. Arai, M. Mori, S. Kurakake, and H. Kojima, “Telop on Demand: Video Structuring and Retrieval based on Text Recognition,” Proc. ICME2000, pp.759-762, 2000.

[13] Y. Taniguchi, A. Akutsu, and Y. Tonomura, “PanoramaExcerpts: Extracting and Packing Panoramas for Video Browsing,” Proc. ACM Multimedia97. pp.427-436, 1997.

[14] 日高浩太, 町口恵美, 竹内順次, 水野理, 中嶌信弥, “音声の感性情報に着目したマルチメディアコンテンツ要約技術,” インタラクション 2003, pp.17-24, Feb.2003.

[15] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura, “Video Handling with Music and Speech Detection,”.Proc. IEEE Multimedia1998, vol.5, no.5. pp.17-25, 1998.

[16] 長田秀信, 紺谷精一, 森本正志, “フィードバックを用いた映像とシナリオ文書の自動対応付け手法,” 信学技報, PRMU2004-53, pp.25-30, 2004.

[17] 三上弾, 紺谷精一, 森本正志, “突発音検出と教師なしクラスタリングを用いた野球映像からの投球イベント検出,” 信学技報, PRMU2004-139, PP.31-36, 2004.