映像・音声認識，言語処理の適用による経済化 メタデータ生成技術

シーン検索，ダイジェスト視聴等の高度視聴サービスが可能な番組映像の流通を促進するために，番組中のシーン内容に関するメタデータの生成作業のコスト削減が重要な課題となっています．映像・音声認識，および言語処理の技術を適用し，作業の低コスト化を実現するメタデータ生成技術を紹介します．

桑野 秀豪（くわの ひでたか） /松尾 義博（まつお よしひろ）
川添 雄彦（かわぞえ かつひこ）
NTTサイバーソリューション研究所

メタデータ生成のコスト問題

番組映像をシーン単位で検索したり，見たいシーンだけをダイジェスト視聴するといった高度な番組視聴サービスを実現するためには，放送前に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ります．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ります．

このような番組の内容に関するメタデータの生成作業をすべて人手で行うと膨大な手間が必要となります．番組制作のコストを下げ，より多くの番組映像を流通させるためには，メタデータの生成作業を効率化することが重要な課題となっています．本稿では，放送通信連携システムにおいて流通するさまざまなメタデータの中でも生成コストがもっともかかる番組内容に関するメタデータを効率的に生成するための映像・音声認識，言語処理技術，および開発したシステムを紹介します．

セグメントメタデータとその生成作業

番組内容を記述するメタデータの国際標準形式がTV-Anytimeフォーラム[(1)]で策定されています．TV-Anytimeでは，番組映像中のシーン区間を「セグメント」と定義し，セグメントを説明するための情報を「セグメントメタデータ」と呼んでいます，セグメントメタデータの内容としては，そのセグメントの番組中における開始時間，終了時間，およびセグメントのタイトル，概要，キーワードといったテキスト情報があります．以降では，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ます．図１にニュース番組についての区間メタデータ，意味メタデータの具体例を示します．

手作業で区間メタデータの生成作業を行う場合，番組映像を最初から最後まですべて再生視聴して内容を確認する必要があります．チェックしたいシーンにたどり着くまでの映像再生，早送りにかかる時間が作業時間の大部分を占めます．早送りしながら見たいシーンを見逃さないように集中する必要があるため，作業者の疲労感も大きい作業になります．また意味メタデータについては，同じく番組映像をすべて再生視聴しながら，番組内容を理解し，適切なタイトル，概要，キーワードといった情報を付与することになります．番組ジャンルやサービス内容にも依存しますが，手作業の場合のセグメントメタデータ生成の作業時間は場合によっては，番組の長さの10倍以上にもなることがあります．このようなことから，セグメントメタデータの生成作業の効率化のための技術が強く求められていました．

映像・音声認識，言語処理の適用による セグメントメタデータの自動生成

セグメントメタデータの生成作業の効率化のためには，特に手作業で行った場合に時間がかかる部分が自動化できると効果的です．

区間メタデータの生成作業には，映像内容の全体の流れが簡単に参照でき，また見たいシーンを確認する際に，番組中の任意の指定したシーンからすぐに映像が再生され内容が確認できると，作業時間の短縮化が見込めま

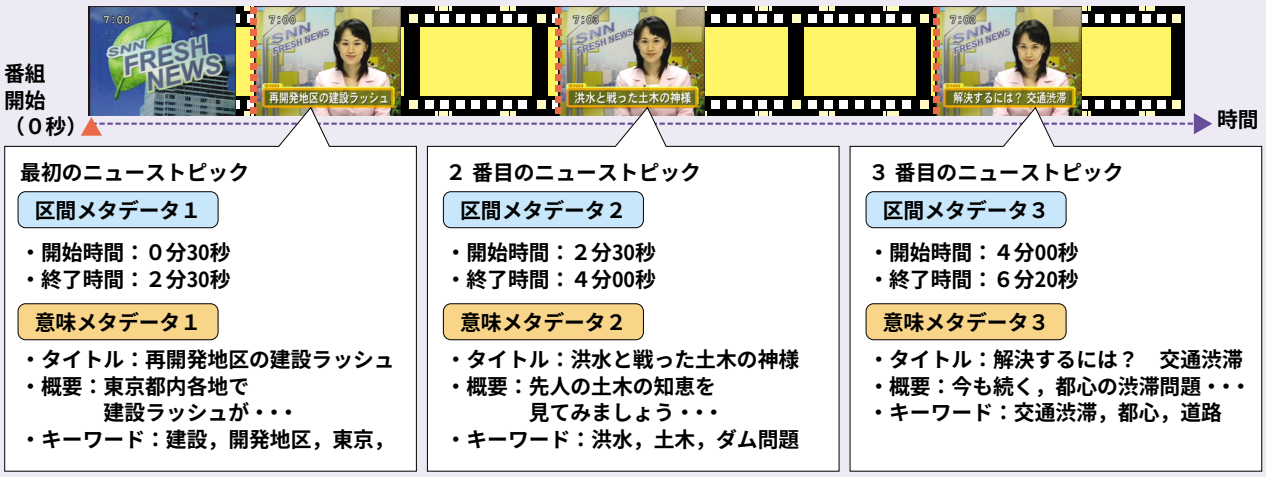


図１　ニュース番組におけるセグメントメタデータの例

す．これには，NTTサイバースペース研究所で開発した映像インデクシングシステムSceneCabinet[(2)]が有効に利用できます．SceneCabinetには，映像認識技術により，映像中のカット点，テロップの文字といった映像内容を確認するうえでポイントとなるイベントを自動抽出する機能，また抽出したイベントを静止画像の一覧表示で見やすく表現し，さらに，画像を選択するとその画像から始まるシーンの再生がすぐに行える機能が備わっています．

意味メタデータの生成作業は，作業者が映像内容を確認してから付与するのではなく，映像内容を確認しなくても必要なテキスト情報が自動的に取得できると作業時間の短縮化が見込めます．これには，前述のSceneCabinetに加え，同じくNTTサイバースペース研究所で開発された音声認識エンジンVoiceRex[(3)]，音声インデクシングシステム[(4)]が有効に利用できます．これらのシステムは，映像中のテロップ文字，音声の内容を自動認識し，テキスト化する機能，さらには，認識結果のテキスト情報から映像のタイトル，概要，キーワードとしてふさわしいセンテンス，単語情報を言語処理により自動抽出する機能を持っています．

これらのセグメントメタデータの自動取得機能によりいったん情報を取得しておき，作業者はその内容を確認し，必要に応じて修正するという作業フローにすることで，セグメントメタデータ生成全体の作業時間の短縮化が可能となります．

経済化メタデータ生成システム

前述の３つのシステムの機能をベースとして「経済化メタデータ生成システム」を開発しました．図２にシステムの機能構成を示します．システム全体としては，「メタデータ自動生成エンジン部」と「メタデータオーサリングGUI（Graphical User Interface）部」から構成されます．

メタデータ自動生成エンジン部は映像中のカット点，カメラワーク，音楽の検出，およびテロップ文字，音声の認識を行い，検出した各シーンの時間情報と画像データ，各認識結果はテキストデータとして出力します．また言語処理により，テロップ文字，音声の認識結果から映像のタイトル，概要，キーワードを自動抽出します．さらには，テロップ文字の誤認識結果を単語辞書を用いて修正する機能も備えています．

メタデータオーサリングGUI部は，メタデータ自動生成エンジン部の処理結果を基に，最終的なセグメントメタデータを完成させるための作業者向けGUI機能を提供します．メタデータ生成の作業効率化には，自動的に生成した情報を確認，編集するユーザインタフェースのデザインも重要なポイントとなります．図３にメタデータオーサリングGUI部の画面例を示します．

メディアブラウザは，「映像インデクシング部」の処理結果の画像データを一覧表示し，映像全体を簡単に一括参照できます．各画像はカット点，テロップ文字など，シーン種別ごとに色分けして表示され，確認したい種別の画像だけを取捨選択することができます．再生モニタは，メディアブラウザ中の画像を選択するとその画像からすぐに映像を再生させることができ，かつ，映像中のシーン画像をコマごとや10秒ごとといったさまざまな粒度の間隔でジャンプさせて確認することがで

きます．メディアブラウザと再生モニタを利用し，区間メタデータを簡単に確認，修正することが可能となっています．さらに，再生モニタには，音声レベルの波形表示機能がついていることもポイントの１つです．一定の時間区間の音声の有無が映像を再生することなく確認できるメリットがあります．

メタデータエディタでは，編集対象の映像区間内に含まれるテロップ文字，音声の認識結果のテキスト情報を基にその映像区間のタイトル，概要，キーワード，代表画像を自動抽出し，その結果を確認・編集できます．

さらに，メタデータオーサリングGUI部では，例えば，ニュース番組における番組進行表，アナウンサーの読み原稿など，番組映像の台本，シナリオといった既存のテキスト情報の電子ファイルをインポートし，これらを基に，意味メタデータを自動取得する機能も備えています．台本，シナリオが利用できる場合には，映像・音声認識処理

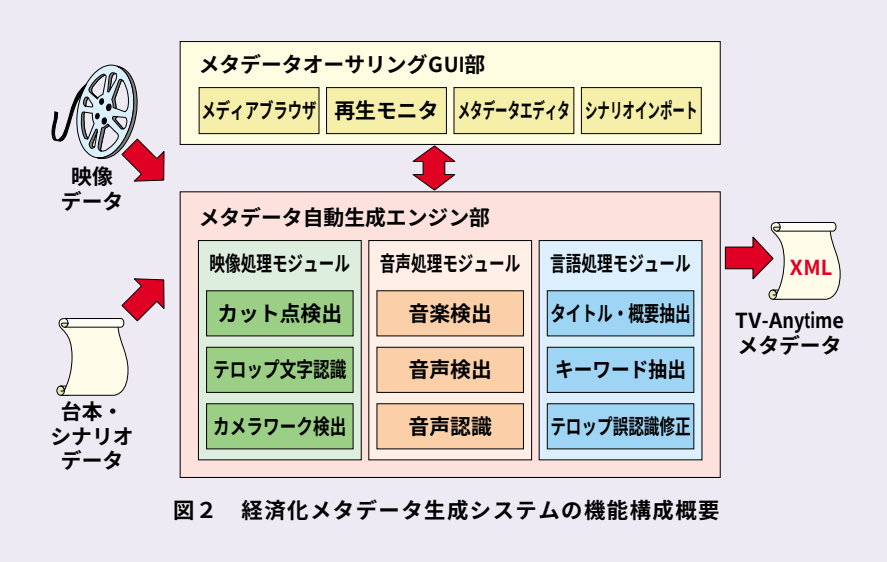


図２　経済化メタデータ生成システムの機能構成概要

図３　メタデータオーサリングGUI部の画面例

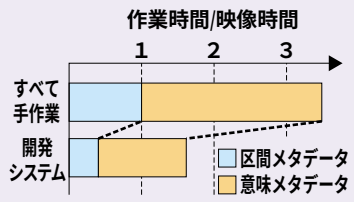


図４　作業モデル別の作業時間

の結果と組み合わせて，より適切な意味メタデータを自動取得することが可能となり，一層の作業時間の短縮化が可能となります．本システムで作成されたセグメントメタデータは最終的にTV-Anytime形式のメタデータファイルとして出力されます．

セグメントメタデータ生成作業の効率化検証実験

ニュース番組を対象に図１で紹介した区間メタデータ（各ニューストピックの開始時間と終了時間）と意味メタデータ（各ニューストピックのタイトル，概要，キーワード）を生成する作業について，メタデータ生成システムを用いた場合とすべて手作業で行った場合の作業時間の比較実験を行いました．複数の作業者により，それぞれの作業モデルごとの作業時間を計測しました（図４）．

区間メタデータ，意味メタデータ，両方の生成作業時間の合計を比較しました．すべて手作業で行った場合は映像時間の約3.5倍，開発システムを用いた場合は約1.6倍でした．このことから開発システムにより作業時間が手作業の場合の約46％に短縮されることが分かりました．作業時間の内訳としては，区間メタデータ，意味メタデータともに作業時間が45％前後に短縮できました．

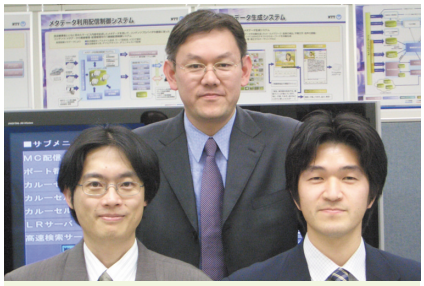
作業者のアンケート結果によれば，区間メタデータの生成には，特にメディアブラウザに表示されるアナウンサーの画像，テロップの画像を手がかりにすると，ニューストピックの開始時間，終了時間を簡単に探すことができたという結果が得られました．さらに再生モニタの各種再生制御ボタン，音声レベルの波形表示機能も作業中の映像再生時間，早送り時間を減らすうえで有効であるとの結果が得られました．また意味メタデータの生成は，メタデータエディタの映像要約機能で各ニューストピックのタイトル，概要，キーワードが適切なものとして取得でき，一部を修正するだけの作業で済んだことから作業時間，および作業による疲労感も人手作業に比べ少なく済むという結果が得られました．

メタデータ生成のさらなる効率化のために

本稿では，セグメントメタデータを用いた番組視聴サービスを実現するために必要なメタデータ生成技術，開発したシステムを紹介しました．ニュース番組の場合，主要な内容がテロップの文字やアナウンサーの発話内容により表現されるため，映像・音声認識，言語処理により，最終的に生成したいメタデータの内容に近い情報を自動的に得ることができ，作業時間の効率化に大きく寄与することが確認できました．今後は，ドラマ，スポーツといった対象番組ジャンル拡大のための映像・音声認識，言語処理の各機能の高度化，および番組ジャンル，生成するメタデータの内容によって変わり得るメタデータ生成の作業フローに即した直感的で簡単に操作できるユーザインタフェースの研究開発に取り組んでいきます．

■参考文献

(1) http://www.tv-anytime.org/
(2) 谷口・南・佐藤・桑野・児島・外村：“SceneCabinet：映像解析技術を統合した映像インデクシングシステム，”信学論，Vol.J84-DII，No.6，pp.1112-1121，2001.
(3) 野田・山口・大附・今村：“音声認識エンジンVoiceRexを開発，”NTT技術ジャーナル，Vol.11，No.12，pp.14-17，1999.
(4) 林・松永・松尾：“音声認識・言語処理の適用によるコンテンツ内容記述メタデータの生成，”NTT技術ジャーナル，Vol.15，No.4，pp.21-24，2003.



（左から）松尾　義博/ 川添　雄彦/
桑野　秀豪

メタデータを利用した便利な番組視聴もメタデータを付けないことには何も始まりません．番組映像の流通促進のため，メタデータ生成のより一層のコスト削減を実現するための研究開発に取り組んでいきます．

◆問い合わせ先
NTTサイバーソリューション研究所
第一推進プロジェクト
TEL　046-859-2508
FAX　046-855-3495
E-mail　kuwano.hidetaka@lab.ntt.co.jp