映像・音声認識・自然言語処理によるメタデータ生成の作業コスト削減に関する研究

映像・音声認識・自然言語処理による
メタデータ生成の作業コスト削減に関する研究

２０１８年　３月

桑野　秀豪

映像・音声認識・自然言語処理による
メタデータ生成の作業コスト削減に関する研究

桑野　秀豪

システム情報工学研究科

筑波大学

２０１８年　３月

概要

本研究は、放送番組を通信ネットワーク上で配信する映像視聴サービスの市場拡大を目的として取り組むものである。通信ネットワークのブロードバンド化や視聴端末の多様化により、番組全体ではなく、番組途中の個々のシーン単位で視聴するという新たな視聴スタイルへの需要が高まっている。例えば、ニュース番組中の個々のニューストピックであり、また、野球中継の一人一人の打者のシーンを検索して視聴する等のイメージである。特に、屋外のモバイル環境においては、家の中にいる場合よりも時間が無いことが多く、シーン単位の映像視聴による時間的なメリットを享受できる機会が多くなる。

こうした視聴スタイルを実現するためには、予め映像視聴サービスの提供側で、番組全体における各シーンの時間情報やシーン内容を説明する情報を生成しておくことが重要である。本研究では、これらの情報をメタデータと呼ぶ。放送番組は社会的に高い信頼を得ていることが多いため、そのメタデータについても、絶対的な信頼度が期待されているという社会的要請がある。このため、メタデータを生成する作業過程では人手による確認が必須とされている。このことが大きな作業コストとなり、その高コストが映像視聴サービス普及の妨げとなりかねない問題となっていた。

本研究では、シーン単位の映像視聴サービスが次々に生み出され、映像配信ビジネス市場が拡大していくべきと考え、その実現に必要なメタデータ生成の作業コストを削減することを技術課題として取り組む。この課題に対し、映像中のテロップ文字の認識をはじめとする、映像・音声認識・自然言語処理等のメディア解析技術により、メタデータを自動生成する方式を提案する。特に、テロップ文字の視覚特徴と番組映像の意味内容の相関に着目し、シーンの時間情報を自動生成することを特徴とする方式である。

また、メディア解析技術により自動生成したメタデータ候補を作業者に分かりやすく提示し、効率的なメタデータ生成作業を可能とするユーザインタフェースシステム「SceneCabinet / NBS」及び「SceneCabinet / Live!」を提案する。これらのシステムを用いて、人手作業の手間が極力少なく済む作業モデルを策定し、制作済番組、及び、ライブ番組に対し、メタデータ生成の作業コストの削減効果を検証する実験を行った。提案方式は、作業を全て人手で行う場合に比べ、作業時間を大幅に削減でき、映像視聴サービスの市場拡大に貢献できるものであることを示す。

目次

第1章 序論

1.1 研究の背景

日本国内での通信ネットワークを活用した一般家庭ユーザ向けの映像視聴サービスは 2000 年頃から本格的に商用サービスが立ち上がった。それ以降、現在に至るまで、視聴されるコンテンツタイトルの数、視聴ユーザ数は増え続けている。同時にコンテンツの視聴スタイルの多様化も進んできた。

2000 年頃、一般家庭向けの通信サービスはメタル回線を利用した ADSL 方式によるブロードバンド化が始まった。当時の通信速度は 1.5Mbps から最大で 50Mbps であった。ブロードバンド上の映像視聴サービスの先駆けとして、2001 年、NTT ブロードバンドイニシアチブ社から、MPEG1 等の映像圧縮が施された映画、テレビドラマ、アニメ等のコンテンツ配信サービスが始まった。

コンテンツ配信の当初から現在まで共通する視聴スタイルとしては、映像コンテンツを先頭から最後まで通して視聴する形が挙げられる。2000 年頃であれば、視聴者はパソコンに搭載されている再生ソフトウェアを利用して、映像コンテンツを先頭から最後まで通して視聴していた。映画等の映像コンテンツを一つの作品として、最初から最後までストーリーをじっくり視聴するこの形は、現在でも視聴スタイルの主流の一つである。

一方で、映像視聴サービスの普及に伴い、ドラマ、映画等の既に完成されたコンテンツを配信する形から、ニュース番組、スポーツや音楽のライブ映像など、リアルタイム性の高いコンテンツまでその適用範囲が広がってきた。この背景としては、2003 年から始まった高速通信サービスが挙げられる。Fiber to the home（FTTH）のコンセプトで、光ファイバ回線による 100Mbps 以上の通信速度を提供するサービスが 2003 年から開始され現在に至っている。これにより、通信ネットワークを介して提供される映像コンテンツの数が増え、また、コンテンツの新たな視聴スタイルを呼び起こすこととなった。

映像コンテンツの新たな視聴スタイルとして、特に、ニュース番組やスポーツ番組では、番組全体ではなく、番組中の個々のニューストピック、あるいは、

野球のホームラン場面、サッカーの得点シーン等、試合中の盛り上がりのあるシーンだけの視聴を求める需要が現れてきた。それに対応する形で、映像視聴サービス上で、該当するシーンを映像コンテンツ全体から切り出して、それらが一つ一つ別々の映像として配信されるようになった。この視聴スタイルにより、視聴者は、長時間の全体映像の中から、見たい場面を探す必要なく効率的に視聴できるメリットが享受できるようになった。

更に、2007 年頃からは、それまでのコンテンツの制作事業者側から視聴者に対する一方向の映像提供に加え、web2.0 というコンセプトに代表されるような、視聴者側から映像内容に対するコメントを発信・共有し、映像と合わせた視聴するスタイルが普及し始めた。主要なサービスとして、ニコニコ動画や YouTube が挙げられるが、これらは、視聴者自身が映像を制作し、発信する映像共有サービスであり、多くの視聴者に利用されている。

また、映像視聴を視聴するデバイス機器の観点からみると、2003 年頃から、それまで主流だったパソコンに加え、通信能付きのセットトップボックスや大画面テレビが発売されると共に、2006 年頃からは携帯電話、また、2008 年の iPhone 発売に始まるスマートフォン爆発的な普及を受け、モバイル映像配信サービスも始まり、家庭内外で広く映像視聴できる機会が増えた。特に、視聴デバイスの多様化は、映像を視聴する機会の増加に貢献した。このことは、同時に視聴者は細切れの時間を映像視聴に当てることが多くなったことを意味している。

こうした状況から、映像視聴サービスの提供者にとって、映像コンテンツのシーン単位への適切な分割と管理が重要な課題となっている。本研究は、通信ネットワーク上での映像コンテンツ配信産業の発展、特に、映像のシーン単位での視聴サービスの普及、市場拡大に貢献することを指向して取り組むものである。次節で、本研究の具体的な目的を述べる。

1.2 研究の目的

本研究では、前項で述べた通信ネットワークを活用した映像視聴サービスの中でも、特に、「シーン単位の映像視聴」に着目し、技術課題の抽出と課題の達成に取り組む。シーン単位の映像視聴の具体例としては、最新のニュース番組から興味のあるニューストピック映像だけを選択して視聴したり、あるいは、野球のナイター中継をリアルタイム視聴はできなかったが、外出からの帰宅後に、リビングのテレビで追っかけダイジェスト視聴したりする等、映像コンテンツ全体の中の見たいシーン映像だけを短時間で視聴することが挙げられる。

他の具体的な例として、過去何十年か分の大量の放送番組映像アーカイブスに対し、シーン単位で視聴できるようインデクスをデータベース化することで、好きな俳優やスポーツ選手だけの映像や、スポーツの得点シーンなどの盛り上がりのシーンを検索して視聴するサービスが挙げられる。

シーン単位の映像視聴は、視聴者にとって、映像全体の中から見たいシーンを探す手間が省け、また、視聴時間の長さも映像全体より短く済むものである。そのため、視聴者は映像視聴の利便性や時間の効率性等の観点で高い満足感が得られる視聴スタイルである。前項で述べた通り、特に、屋外のモバイル環境においては、家の中にいる場合よりも時間が無いことが多く、シーン単位の映像視聴による時間的なメリットを享受できる機会が多くなる。

こうした視聴スタイルを実現するためには、予め映像視聴サービス提供側で、映像コンテンツを適切にシーン分割するための各シーンの時間情報、及び、そのシーンを適切に記述する情報、本研究では、これら情報をメタデータと呼ぶが、このメタデータを映像視聴サービスの提供前に生成しておくことが重要である。特に、放送番組のような映像コンテンツは社会的に高い信頼を得ていることが多いため、そのメタデータについても、絶対的な信頼度が期待されているという社会的要請がある。このため、その作業過程では人手による確認が必須とされている。このことが大きな作業コストとなり、その高コストが映像視聴サービス普及の妨げとなりかねない問題となっていた。

本研究では、通信ネットワーク上での映像コンテンツ配信産業において、シーン視聴による高いユーザ利便性を提供するために必要な、映像コンテンツのシーン単位への適切な分割と管理を産業界の望む形で効率化する方法について述べる。そのために、映像処理、音声認識、言語処理等の映像コンテンツの解析と、そこからの人手による確認を含んだ効率の良いメタデータ生成法に取り組む。

1.3 研究の課題

本研究は、放送番組等、多額の費用をかけて制作される映像コンテンツを対象に、シーン単位の映像視聴サービスを提供する事業者の立場における技術課題を設定し、取り組むものである。

放送番組の提供サービスは、前項で述べた社会的信頼と共に、一般に、サービス要件に対し、番組制作のスポンサー企業の意向の影響が強いことが多い。

すなわち、シーン単位の視聴サービスにおいても、スポンサー企業の事業推進上、不適切な情報が提供されてはいけない等の厳格なサービス要件が存在する。

1 本の放送番組から、シーン単位の映像を制作する際には、番組映像コンテンツ全体の中のどのあたりの時間にどのようなシーンがあるのかという、いわゆる、映像内容全体の中の 1 つ 1 つのシーンに関する目次、インデクスにあたる情報を生成する必要がある。社会的信頼やスポンサー企業の意向を考慮しながら、1 つ 1 つのシーン自体が、1 つの作品として、一定の情報クオリティを満足した形で提供される必要がある。例えば、ニュース番組中の 1 つ 1 つのニューストピックのシーンに関する時間や内容に関する情報が該当する。このように、放送番組を対象とする際には、一般ユーザによる投稿動画のような低コスト映像のサービスとは一線を画すサービスクオリティが求められる。

前項でも述べたが、本研究では、この番組映像内容に関する目次、インデクスにあたる情報を「メタデータ」と呼ぶ。メタデータは、前述の通り、厳格なサービス要件を満たす必要があることから、それを作り出すには、人間の目、耳、手作業を駆使して、内容チェックしながら作り出す必要がある。しかしながら、メタデータの生成を全て人手作業で行うと、映像内容を膨大な作業時間等のコストがかかる。このことは、シーン単位の映像視聴サービスという、大きなビジネスポテンシャルを持つサービスの普及、浸透を妨げることとなり、大きな社会的損失となっていた。

本研究では、シーン単位の映像視聴サービスが次々に生み出され、通信ネットワーク上での映像配信ビジネス市場が拡大していくべきと考え、その実現に必要なメタデータ生成の作業コストを削減することを技術課題として考える。この技術課題に対し、本研究は、メタデータ生成の作業コストを削減するための方式を実装したシステム「SceneCabinet / NBS」を提案する。SceneCabinet / NBS を用いたメタデータ生成の作業モデルでは、最初に、映像・音声・自然言語といった映像中のメディア内容を認識し、区間メタデータ、意味メタデータの手がかりとなる情報を取得し、人手作業で確認、修正を行う。

特に、メディア認識では、映像中のテロップ文字の特徴に着目した方式を提案する。人手作業では、作業を最小限の時間で完了することを指向としたユーザインタフェースシステムを提案する。SceneCabinet / NBS は、これらメディア認識機能とユーザインタフェース機能を備えたシステムである。本研究では、SceneCabinet / NBS を用いて、制作済の番組映像、また、SceneCabinet / NBS をベースとしたシステム SceneCabinet / Live! を提案、活用し、ライブ放送の番組映像に対して、メタデータ生成作業の時間短縮効果を評価し、提案方式の有効性を明らかにする。

1.4 論文の構成

本論文の構成は以下の通りである。図 1.4-1 に全体構成の俯瞰図と筆者らの公表論文との対応関係について示す。

なお、本論文の内容は、2000 年から 2007 年の研究成果をまとめたものである。現在は、研究当時から約 10 年が経っており、最新の技術動向や映像サービスの動向は、研究当時から大きく変わっている部分もある。本論文内では、研究当時の提案内容や実験結果、産業貢献の内容等を中心に論じるが、各章のまとめや論文全体の結論においては、最新の映像サービス、技術動向に対する本研究の貢献内容についても合わせて整理して述べる。

第 1 章　序論

本研究テーマ「映像・音声認識・言語処理によるメタデータ生成の作業コスト削減に関する研究」の背景、目的、及び、課題を明確にし、本論文の構成を示す。

第 2 章　メタデータ生成の作業コストに関する課題

映像コンテンツに対するメタデータ生成の作業コストに関する課題を整理して述べる。映像コンテンツのシーン視聴サービスの実現にあたって、メタデータの必要性、メタデータの定義、及び、メタデータ生成に関する要件を述べ、次いで、従来行われているメタデータ生成の具体的な作業内容とコストに関する課題を述べる。メタデータ生成の作業については、放送事業者における番組映像の制作作業フロー全体の中における位置づけを整理して述べ、その観点からの作業コストに関する課題を述べる。映像制作の作業フローは研究当時のアナログベースから、最新の 2018 年現在では完全デジタル化という大きな変化を遂げようとしている。この変化と本研究での課題の関係性も整理して述べる。

第 3 章　メタデータ生成の作業コスト削減に向けたアプローチ

メタデータ生成の作業コスト削減に向け、本研究が提案する方式のアプローチを述べる。最初に、方式提案にあたっての基本的な考え方を整理して述べる。放送番組映像に対するメタデータ生成に対し、「人手作業の必要性」「作業コストの内容と削減の可能性」「メタデータ生成の効果的な自動化」「ユーザインタフェースのデザイン」といった観点から分析を行う。次いで、提案方式の策定にあたっての考え方を述べる。大きく 2 つのステップから構成することとし、第 1 ステップとして、メタデータの手がかりとなる情報をメディア解析技術で自動的生成する。メディアの中でも、特に、番組映像中のテロップ文字と映像の意味内容の相関に着目するアプローチを取る。第 2 ステップとして、その結果を人手作業で確認、修正し、メタデータを完成させる。特に、作業を短時間化するためのユーザインタフェースを活用するアプローチを取る。

第 4 章　テロップ文字認識によるメタデータ自動生成

メタデータ生成の作業コスト削減の方式提案と評価、考察は第 5 章、第 6 章で述べるが、本章では、人手作業の事前に行う処理であり、かつ、作業コスト削減のためのブレークスルーとなる方式として、映像中のテロップ文字の認識式を述べる。特に、従来にない新しいアプローチとして、番組映像中のテロップ文字の持つ文字の大きさや画像中の表示位置等の視覚的特徴と映像の意味内容には相関性に着目した方式を提案する。例えば、ニュース番組中のニュートピックの冒頭シーンや野球中継の得点シーンでは、必ず、決まった大きさで、決まった位置にテロップ文字が表示される。提案方式に加え、実映像データによる提案方式の評価結果と考察を述べる。

本章の内容は、筆者らの公表論文[C1],[C2],[C3]に基づくものである。

第 5 章　制作済番組向けメタデータ生成の作業コスト削減

放送した番組映像をネット配信向けに 2 次利用するケース、すなわち、制作済番組を対象とし、メタデータ生成の作業コストの削減方法の提案、実験結果、及び、考察を述べる。第 3 章で述べるステップに基づき、第 4 章で述べるメデ

ィア解析技術によるメタデータの手がかり情報を生成する機能、また、その内容を人手作業でチェックし、メタデータを完成させるユーザインタフェース機能を備えたシステム「SceneCabinet / NBS」を提案する。SceneCabinet / NBS を活用したメタデータ生成の作業モデルを検討、提案し、制作済のニュース番組、サッカー中継番組に対するメタデータ生成の作業コスト削減効果の評価実験の結果と考察を述べる。これらを踏まえ、研究当時、及び、最新の映像サービス、技術動向に対する本研究の貢献内容を明らかにする。

本章の内容は、筆者らの公表論文[J1],[C4]に基づくものである。

第 6 章　ライブ番組向けメタデータ生成の作業コスト削減

メタデータを利用した番組視聴サービスの適用範囲の拡大を指向し、ライブ放送される番組を対象として、番組放送中にリアルタイムにメタデータを生成する作業モデルを提案し、人手作業との比較実験の結果と考察を述べる。番組の放送中に、番組進行に遅れることなく、メタデータを次々に生成するという課題に対し、前章で述べる「SceneCabinet / NBS」をライブ放送用に拡張したユーザインタフェースシステム「SceneCabinet / Live!」を提案する。SceneCabinet / Live! は、メタデータ生成の作業者が、音声で情報入力するだけで、メタデータ生成可能なユーザインタフェースを提供する。ライブ野球中継に対し、SceneCabinet / Live! を活用したメタデータ生成の作業モデルを適用する実験結果、及び、考察を述べる。

本章の内容は、筆者らの公表論文[J2],[C5]に基づくものである。

第 7 章　結論

メタデータ生成の作業コストの削減という課題に対し、映像、音声認識、自然言語処理を活用し、また、人手作業向けのユーザインタフェースを利用する提案方式についての到達点、及び、映像視聴サービス産業にもたらす貢献について考察を行う。研究当時の産業貢献と、最新の関連サービス動向、関連技術動向に対する貢献を整理して述べる。

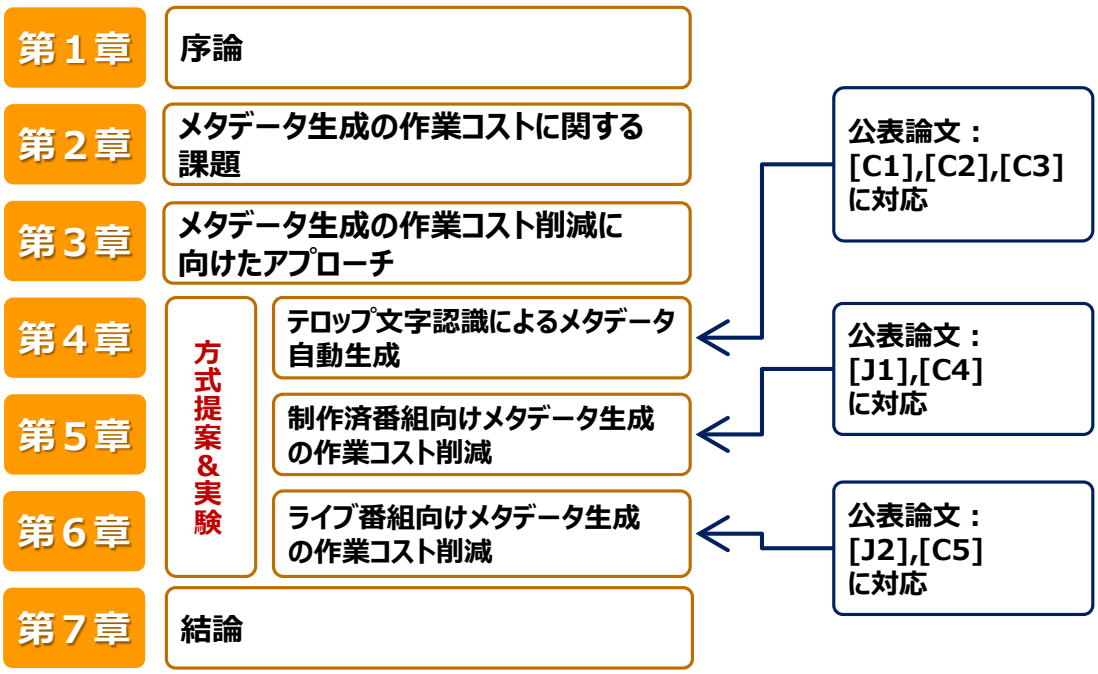


図 1.4-1　本論文の構成と公表論文との関係性

第2章 メタデータ生成の作業コストに関する課題

2.1 概要

本章では、映像コンテンツに対するメタデータ生成の作業コストの削減にあたっての各種動向と具体的な課題を整理して述べる。最初に、放送番組映像のシーン視聴サービスの動向について述べる。次いで、同サービスの実現にあたって、メタデータの必要性、メタデータの定義、及び、メタデータ生成に対する要件を順に述べた後、従来のメタデータ生成の作業内容とコストに関する課題を述べる。メタデータ生成の作業は、一般に、放送事業者、ネット配信事業者における番組映像の制作作業の中の一つとして実行される。この番組映像の制作作業全体の観点も含め、メタデータ生成の作業コストに関する具体的な課題を述べる。

なお、放送事業者等における映像制作の作業フローは、本研究当時は、映像コンテンツをアナログベースで扱うことが主流であった。しかしながら、近年、2018 年においては、ファイルベースとなり、完全デジタル化という大きな変化を遂げようとしている。この変化と本研究の課題の関係性も整理して述べる。

2.2 番組映像のシーン視聴サービス

1.1 節「研究の背景」で述べたように、2003 年頃から、テレビ番組や映画等の映像・音声コンテンツを通信ネットワーク経由でパソコン向けに配信するサービスが始まった。その中でも、特に、ニュースやスポーツの試合の番組映像は、放送事業者の情報配信サイトにて、ニュース番組全体の中の個々のニューストピック、野球の試合全体の中の各打者のシーン等、1 つの番組全体の中の、

更に 1 つ 1 つのシーン単位で検索してから視聴したり、あるいは、盛り上がりのシーンを繋げて、ダイジェストの形で視聴したりする等、多様な視聴スタイルの形で発展してきた。

図 2.2-1 は 2017 年 10 月 21 日の TBS のニュース映像配信サイトの画面例である。TBS 系列の複数のニュース番組を対象とし、各ニュース番組が 1 つ 1 つのニュートピック毎の映像シーンに分割され、一つのコンテンツとして提供されている例である。ニュース番組に関しては、日本全国の放送事業者のニュース配信サイトの他、世界的にみても、米国の放送事業者の三大ネットワークや CNN、また、英国 BBC など、ニュートピック単位の映像視聴サービスが提供されている。

図 2.2-1　TBS のニュース映像配信サイトの画面例

（参照元：TBS ニュース映像配信サイト, http://news.tbs.co.jp/）

図 2.2-2 は、スポーツ情報の配信サイト「スポーツナビ」の画面例である。プロ野球の試合中の盛り上がりのシーン映像が一覧、検索できるサービスであり、図中では、ホームランシーンやダイジェストシーンの画像一覧とその内容を説明するテキスト情報が簡単に閲覧できる様子を示している。

このようなシーン単位の視聴サービスは、短時間で映像コンテンツの内容を把握することができ、従来の家庭内でのテレビやパソコンでの映像視聴スタイルに加え、外出先でスマートフォンの画面で視聴するスタイル等、視聴者の生活の様々なシーンでの映像視聴機会を創出する。視聴者の持つ、様々な趣味、嗜好に応えるべく、少しでも多くのシーン映像が流通することが望まれる。このことは、国内外の放送事業者、ネット配信事業者より、大きなビジネスチャンスとして捉えられている。

図 2.2-2　スポーツナビの画面例

（参照元：スポーツナビ情報配信サイト, https://sports.yahoo.co.jp/）

2.3　メタデータの必要性

前節で述べた、シーン単位の映像視聴サービスの実現に必要なメタデータについて、その必要性を述べる。30 分や 1 時間、2 時間の長さの映像全体を見るのではなく、一部のシーンだけを視聴する方法としては、例えば、ハードディスクレコーダーに録画された映像を倍速再生したり、適宜、30 秒ジャンプ再生をしたりすることでも良い場合もある。これは、映像コンテンツをユーザ個人が私的利用の範囲で視聴する場合に考えられる方法である。

このような私的な利用方法ではなく、図 2.2-1、図 2.2-2 に示したように、放送番組をシーン単位の映像に区切り、それぞれを 1 つ 1 つの映像作品として提供する場合、一般には、番組制作のスポンサーの意向を確認しつつ、1 つ 1 つのシーン映像に対し、タイトル等のテキスト情報を付けて、一定の情報クオリティを保証した形で提供されることが多い。例えば、図 2.2-1 のニュース配信サイトの場合、図 2.3-1 に示すように、ニュース番組全体における該当ニュース映像の区切り、すなわち、開始タイミング及び、各ニューストピックのタイトル情報が、サービス提供の事前に一定のクオリティで作り上げられている必要がある。

図 2.3-1　ニュース配信サイトにおけるメタデータの利用例

番組映像に対し、1 つ 1 つのシーンの区切りや、シーンのタイトル情報などは、シーン映像を説明するための情報（データのためのデータ）として、「メタデータ」と呼ぶ。番組映像全体に対し、メタデータを生成しておくことで、メタデータの一覧や検索により、効率的に番組映像中の見たいシーン映像を視聴することができる。次節では、メタデータの定義について述べる。

2.4　メタデータの定義

前節で述べたメタデータは、映像の個々のシーン内容を説明する情報であり、書籍でいうところの目次や索引にあたる。例を挙げると、ニュース、紀行番組、料理番組といった情報提供型の番組や、野球、サッカーといったスポーツ中継番組において、各ニュースのトピックや、あるいは、試合におけるシュートのシーン、ホームランのシーン等が、番組全体の中のどのあたりの時間で起こったかという時間情報、及び、シーンの内容を説明するテキストが具体的なメタデータの項目になる。本研究では、メタデータのうち、シーンの時間情報を区間メタデータ、また、シーンのタイトル、概要、キーワードといったテキスト情報を意味メタデータと呼ぶこととする。

図 2.4-1 に、ニュース番組における、区間メタデータ、意味メタデータの具体例を示す。ニュース番組は、通常の地上波放送では、複数のニューストピックが順番に放送されるという構成をとる。ニュース番組の場合、ニューストピック単位での視聴サービスを実現するため、「ニュース番組全体における個々のニューストピックの開始時間、終了時間の情報」が区間メタデータであり、トピック単位でのキーワード検索時に必要な「個々のニューストピックのタイトル、概要、キーワード情報」といったテキスト情報が意味メタデータとなる。

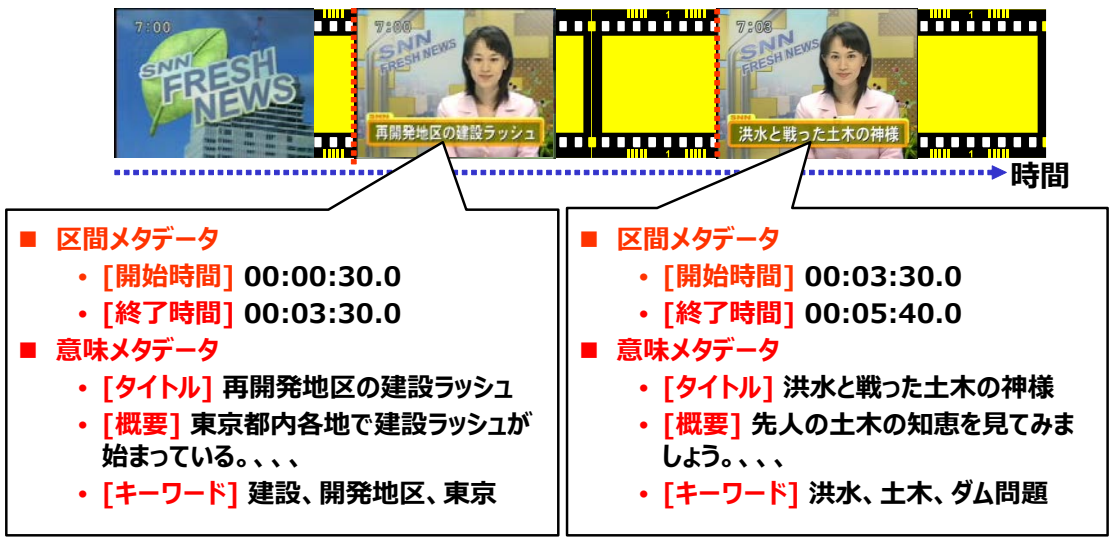


図 2.4-1：　ニュース番組における区間メタデータ、意味メタデータの例

図 2.4-2 には野球番組におけす区間メタデータ、意味メタデータの例を示す。野球番組のシーン視聴としては、一人一人のバッターの打席シーンが考えられる。図 2.4-2 は区間メタデータとして、各打者の打席シーンの開始と終了時間、意味メタデータとして、打席シーンの内容を示すタイトル、概要、キーワードとなる。

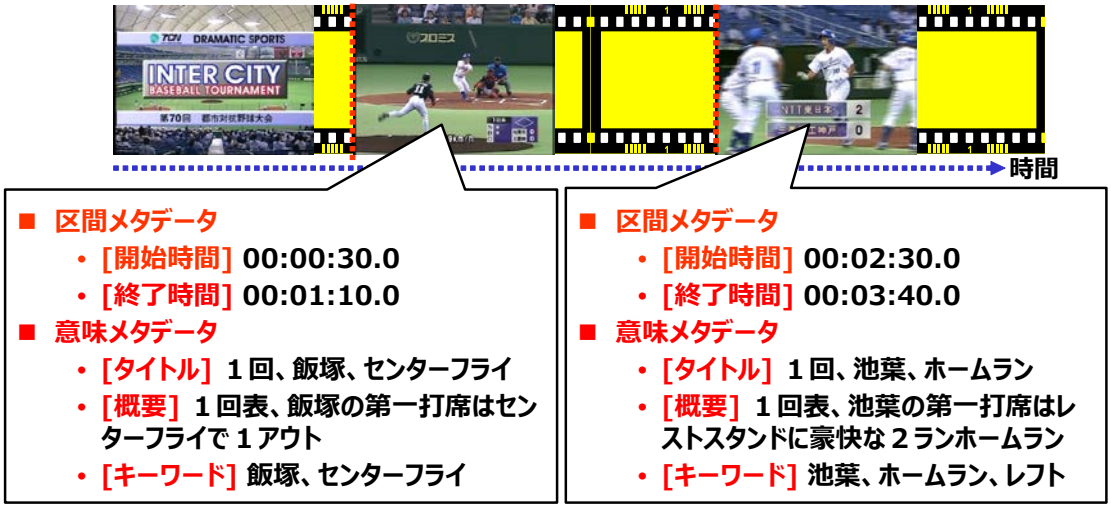


図 2.4-2： 野球番組における区間メタデータ、意味メタデータの例

このように、メタデータは、個々のシーン映像に対し、作品としての価値を与える情報であり、放送番組のシーン視聴サービスの実現にあたって、非常に重要でキーとなる。メタデータは、シーン視聴サービスで、サービス提供者から利用者への流通や、サービス提供者間での流通などを想定し、そのフォーマットの共通化に関する活動も進んでいる。

具体的には、通信放送連携サービス仕様に関する国際的業界フォーラム TV-Anytime Forum で、メタデータの項目、フォーマットに関する標準仕様が策定され[TV-Anytime]、欧州通信標準化機構 ETSI で国際標準として採用された仕様がある[ETSI]。同じ仕様が、国内では電波産業会 ARIB にて、技術資料（TR-B38）として策定されている[ARIB]。

いずれも、XML の形式で定義されるものであり、図 2.4-3 にメタデータ全体のうち、シーン情報を記載する区間メタデータ、意味メタデータを抜粋した部分を示す。シーンに関するメタデータは、“SegmentInformation”タグとして定義され、また、その中でも、区間メタデータは、映像全体の中のシーン区間のの開始時間と終了時間を示す“SegmentLocator/StartTime”タグ、”SegmentLocator/EndTime”タグとして定義されている。また、意味メタデ

ータは、シーンのタイトル、概要、キーワードを示す、“Description/Title, Description/Synopsis, Description/Keyword”タグに相当する。2018 年現在、日本国内では、NTT ぷらら社のひかり TV サービス等で、国際標準規格に乗っ取ったメタデータフォーマットが活用されている。

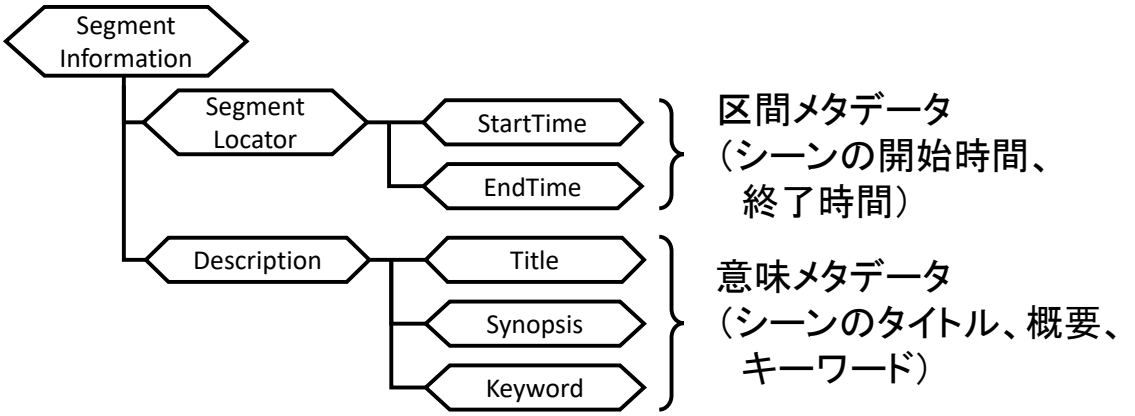


図 2.4-3： TV-Anytime 規格のメタデータ体系（抜粋）と本研究におけるメタデータの関係

2.5 メタデータ生成の作業コスト問題

放送番組映像のシーン視聴サービスの実現にあたっては、前節で定義を述べた区間メタデータと意味メタデータをサービスの事前に生成しなければいけない。本節では、放送事業者、ネット配信事業者におけるメタデータ生成の業務作業について、本研究当時の 2007 年頃における作業内容と課題、また、その後の最新のメタデータ生成の方法を述べ、最新動向に対する本研究の課題設定の位置づけも整理して述べる。

最初に、2.5.1 項で、放送事業者における、本研究当時のメタデータ生成作業として、アナログテープを起点とする番組制作の作業フロー全体、及び、フローにおけるメタデータ生成のプロセスや作業コストに関する課題を述べる。次に、2.5.2 項で、本研究当時に、作業コスト課題に対し、提案されていた研究レベルの取り組みと到達点を整理して述べる。2.5.3 項では、放送事業者における最新の番組制作方式として、デジタルファイルベースの作業フローと、本研究の課題設定の関係性について整理して述べる。

2.5.1　本研究当時のメタデータ生成のコスト問題

2007 年頃の本研究当時の放送事業者における、ニュース番組のネット配信に関する作業フローは、図 2.5.1-1 に示すように、まず、ネット配信ではなく、放送サービス向けに番組制作を行い、放送番組として送出した後に、ネット配信サービス向けに、改めて映像のデジタルファイル化を行うところから行う。すなわち、放送向けとネット配信向けの業務の独立性が高く、放送向けの業務の後に、続けて、ネット配信向けの業務を行うという流れである。

放送サービス向けの番組制作では、アナログテープベースでのニュース取材から始まり、編集後、ニュース番組として送出する。送出後、番組映像をデジタルファイル化し、メタデータ生成の作業者により、当該ファイルをニューストピック毎に分割し、タイトル等のテキスト情報を付与するメタデータ生成作業を行い、デジタル映像ファイル、及び、メタデータ共に、通信ネットワーク上のサーバシステムに格納され、配信される流れである。

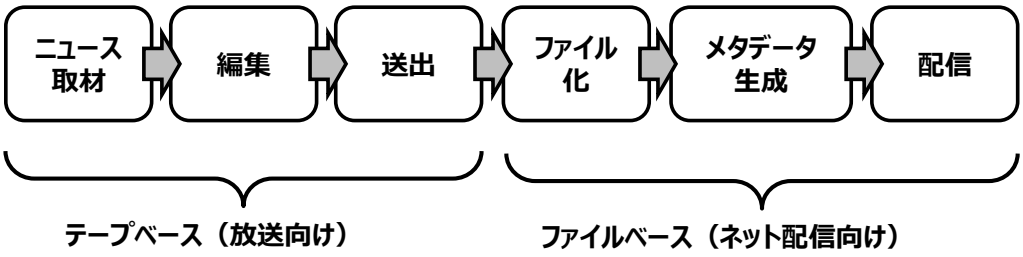


図 2.5.1-1　番組制作のワークフローイメージ

メタデータ生成の作業においては、各ニューストピックのシーンに関する区間メタデータ、意味メタデータを生成するが、これは、従来、放送事業者、あるいは、メタデータ生成の専門事業者により人手作業により行われている。具体的には、専門の作業者がビデオデッキ、ハードディスクレコーダーやパソコンを操作し、映像の再生、早送り、巻き戻し等を行い、目視等で各ニューストピックの始まり、終わりの時間を見つけメタデータとして記録する。また、映像内容を見ながら、あるいは、紙に書かれた番組進行表、アナウンサーの読み原稿といった関連情報を見ながら、タイトル、概要、キーワードを全て手入力する作業である。

メタデータ生成の専門事業者として、エム・データ社が挙げられる。図 2.5.1-2 にエム・データ社での作業現場の様子を示す。エム・データ社では、特定の映像視聴サービスの要件に合わせたメタデータ生成というよりは、様々な映像視聴サービスに共通的に利用することを想定したメタデータ生成を行っている。すなわち、「番組内容の意図を的確に伝える」べく、番組が伝えたい意図をエム・データ独自に、区間メタデータと意味メタデータを生成している。

エム・データ社が生成する意味メタデータは、客観的に要点のみにフォーカスして記述されたテキストであり、「いつ」「どの局、どの番組で」「何が」「誰によって」「どのように」「どの位（時間）」放送されたのかを簡潔に把握することができるものである。このように意味メタデータを生成するためのルールを明確にすることで、メタデータ生成の作業者の違いによる表記の揺れを少なくする工夫がなされている。

図 2.5.1-2 エム・データ社におけるメタデータ生成の現場

以上が、研究当時のメタデータ生成の方法であるが、前述の通り、人手作業により、ビデオデッキ、ハードディスクレコーダーやパソコンを操作し、映像の再生、早送り、巻き戻し等を行い、目視等で各ニュースストピックの始まり、終わりの時間を見つけメタデータとして記録する。また、映像を再生し、出演者の音声内容やテロップ文字表示を把握し、テキスト情報をキーボード等でタイプ入力するものであり、番組映像のジャンルや時間長によって変動しうるが、その作業時間は番組映像の時間長の数倍～10 倍以上かかることもある。

この作業と、放送サービス後のアナログテープ媒体の番組映像をデジタルファイル化する作業にも一定の作業量が必要であり、特に、ニュース番組の場合、番組としての放送後、なるべく間をおかず短時間でネット配信するというサービス要件を満足しようとすると、作業者の疲労感は非常に大きいものであった。

メタデータにより実現される、放送番組のシーン単位の映像サービスは、国内外の多数の放送事業者、ネット配信事業者等から、新たなビジネスチャンスとしての期待が高いものであるが、メタデータ生成の作業時間や作業者の疲労感等、作業環境に関するコスト問題から、メタデータ生成を大きな事業として取り組めてはいなかった。このことは、ユーザにとって、番組映像コンテンツから得られる様々な情報、知識の流通の可能性を制限し、大きな社会的損失を招いていた。このため、なるべく、作業時間や作業者の人数、あるいは、作業に必要な環境が最低限、すなわち、少ない作業コストでメタデータを生成する方法の確立が重要な課題として、国内外の多くの関連事業者から望まれていた。次節は、メタデータ生成の作業コスト削減に関する従来研究とその問題点を整理して述べる。

2.5.2　メタデータ生成の作業コスト削減に関する従来技術

前項 2.5.1 で述べたメタデータ生成のコスト削減に関する課題に対し、従来、関連する技術が研究レベル、製品レベルで存在する。研究レベルでは、国内外の放送事業者や通信事業者、大学等の研究機関を中心に関連の取り組みが行われていた。以下、代表例を述べる。

Macaky らは、デジタル映像に対し、アノテーション（注釈）としてテキスト情報を付与する専用エディターを提案した[Mackay],[Mackay2]。メタデータという概念が生まれる前の 1989 年の研究であり、アナログベースでの取り扱いが主流であった映像コンテンツをコンピュータにデジタル情報としてキャプチャし、映像中の指定したシーンに対し、タイムスタンプやアノテーションを手入力できるユーザインタフェース方式の提案である。映像コンテンツに含まれる、多様な意味解釈ができるマルチメディア情報が、社会学等の様々な分野における資料として有用であるとの期待が高まっていることを捉え、これと、コンピュータのパーソナル化、映像のデジタル化の技術潮流を合わせた研究である。本研究の課題である、メタデータの生成作業コストを削減するレベルまでの検討はなされていないが、映像コンテンツに対し、シーンレベルでテキスト情報を付与する方式研究の黎明期の代表例といえる。

その後、ハードディスクの大容量化が進み、パソコン上でデジタル映像が扱いやすい環境が整い始めた。研究のトレンドとしては、映像コンテンツへのアノテーションに加え、大量の映像コンテンツをアーカイブ化することを念頭におき、メディア解析技術により、映像のシーン分割や索引付け、あるいは、映像内容を代表するサムネイル画像を自動生成する取り組みが多数行われた[Ueda],[Nagasaka],[Hjelsvold],[Kanade],[Satoh],[Nitta]。映像中のカット点、指定した物体、人物の顔等を自動抽出したり、あるいは、字幕テキストを活用したりする方法等、人手による映像アノテーション作業を軽減することを狙った取り組みである。しかしながら、メディア解析技術そのものの精度評価やメディア解析の結果の提示方法の提案が中心であり、本研究の課題であるメタデータ相当の情報を生成する作業コストの削減効果を評価する取り組みはなかった。

その中で、谷口らは、映像中のカット点やテロップ文字が表示されている画像が一覧できる「映像カタログ」を生成する作業の短時間化を目的とした研究を行っている[Taniguchi]。映像中のカット点、テロップ文字等を自動検出し、その結果の静止画像の一覧、編集が実施できるユーザインタフェースシステムを提案している。具体的な作業内容としては、カット点、テロップ文字等の自動検出結果に対し、映像再生機能等のユーザインタフェースを使用しながら、誤検出を取り除く作業である。実験により、同システムによる作業時間は人手作業の約半分の時間で必要な情報が作り出せることを示している。

また、住吉らは、「メタデータ生成に関するフレームワーク」を提案している[Sumiyoshi]。メタデータのフォーマットの国際標準仕様 MPEG7[MPEG]のフォーマットに合わせたメタデータを作り出すことをコンセプトとして、メタデータの手がかりとなる情報をメディア解析モジュールを活用して実施すること、また、その結果を編集することでメタデータを作り出すプロセス全体を整理し、ツールとして実装している。メタデータ生成の作業がどの程度コスト削減できるかの評価に関する情報は公開されておらず、フレームワーク、ツールの提案レベルである。

他にも、メタデータ生成の作業コストに関する具体的な評価結果までは公開されていないが、同様のことを目的とした製品として、[Mediasite]がある。[Mediasite]も[Taniguchi],[Sumiyoshi]と同様、メタデータの生成にあたり、一度、事前に番組映像コンテンツからメタデータの手がかりとなる情報をメディア解析技術で抽出し、その結果を作業者がメタデータ向け GUI を活用して修正を行うモデルである。

しかしながら、従来のいずれの取り組みにおいても、本研究がターゲットと

して考えるニュース番組中の個々のニューストピック、野球中継の得点シーンなど、意味レベルのメタデータを作り出す作業内容を設定し、コスト削減効果の評価までは行われていなかったといえる。図 2.5.2-1 に示すが、[Taniguchi]の取り組みは、映像全体の中の全てのカット点、テロップ文字情報を取り出すものであるが、これは、本研究のニューストピックという意味レベルのメタデータに比べ、より信号レベルに近い情報といえる。

図 2.5.2-1　本研究と従来研究[Taniguchi]における生成対象のメタデータ

以上、2.5.1 項、2.5.2 項は、本研究の実施当時 2007 年頃の番組映像に対するメタデータ生成の作業コストに関する課題と、課題に対する従来の関連する研究、製品を述べた。次項では、2007 年以降の番組映像制作の作業フローと、本研究当時のメタデータ生成の作業コストに関する課題の位置づけについて整理して述べる。

2.5.3　番組映像制作の最新動向

放送番組の映像制作は､本研究当時の 2007 年から現在の 2018 年までの間で、大きな変化、発展を遂げようとしている。2007 年当時は、図 2.5.1-1 に示したように、ニュース番組の制作においては、アナログテープが起点となる取材から始まり、途中、アナログ映像からデジタルファイル化し、メタデータ生成、映像配信という流れであった。

これに対し、2003 年頃、放送番組映像に関し、日本初の大規模な過去映像の公開ライブラリセンターとなる「NHK アーカイブ（埼玉県川口市)」が準備・開業され、その当時から、取材時から映像コンテンツをデジタルファイルとして制作し、その後の全ての工程も一貫してデジタルファイルで行う「ファイルベース」の実現に向けた検討が始まった。ファイルベースは、映像コンテンツが素材ごとに、最初からファイル化され、ランダムアクセスができる等、編集作業等の効率が、テープベースに比べ、飛躍的に向上することについて、期待が非常に高いものであった。

しかしながら、放送業務において、従来から慣れ親しんでいるテープベースのワークフローをファイルベースに移行することは、業務スペース、資金、人員などのリソース計画、移行時の運用ルール等の多数の課題がある他、これを毎日ほぼ 24 時間、放送サービスを提供し続ける中で実行することは、ほぼ不可能と捉えられていた状況が続いていた。

一方で、2010 年頃から、放送事業者の技術開発を担当する部門を中心に、ファイル化の効果が高いと考えられる、アーカイブ設備、次に、カメラ設備、その後、ニュース報道、スポーツ系の各制作部門に少しずつファイルベースのワークフローの取り入れが始まった。また、2015 年には、放送機器メーカーよりテープベースの設備の保守終了タイミングがアナウンスされた。これにより、2018 年時点では、各放送事業者において、番組制作業務の完全ファイルベース化の計画の具体化が進んでいる状況である[Horibuchi]。

図 2.5.3-1 に、従来のテープベースの映像制作フローと、最新のファイルベースのフローを対比して示す。ファイルベースになることで、図 2.5.3-1 下段に示すように、映像コンテンツが、取材時という最初からデジタルファイルとなる。ネット配信用に別途ファイル化をする必要が無くなる効果が最も大きいが、加えて、放送用の編集作業の情報がそのままネット配信用のメタデータ生成に活用できるようになり、テープベースに比べて、作業量が格段に少なくなることが期待される。

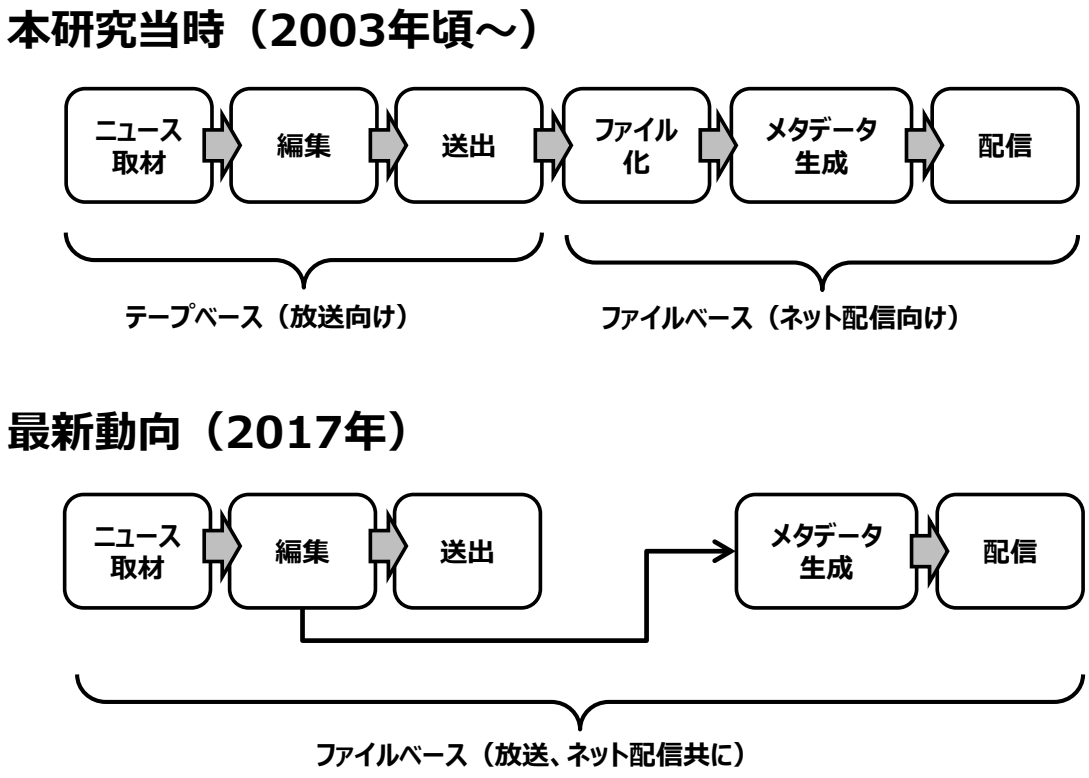


図 2.5.3-1　番組制作のフロー（本研究当時と最新動向）

2.5.4　最新動向における本研究の位置づけ

ファイルベースの番組制作は、取材段階から映像コンテンツがデジタル化されており、取材時の素材映像を放送番組として制作する際の編集作業で生じる様々な情報も全てデジタル情報となる。例えば、素材映像の撮影日時や場所、撮影場所、取材者問い合わせ先の他、編集時に加工されるテロップ文字のテキ

スト情報も最初からデジタル化されて管理される。

放送サービスを番組映像の 1 次利用とすると、本研究がターゲットとする番組映像のネット配信サービスは 2 次利用といえる。2 次利用向けの区間メタデータ、意味メタデータの生成を考える際には、コスト面から、1 次利用時に作られ、既にデジタル情報として存在するメタデータで再利用可能な情報があれば積極的に利用することが望ましい。区間メタデータ、意味メタデータの生成に利用可能なデジタル情報の例として、以下が考えられる。

■　メタデータの生成に利用可能なデジタル情報の例

- ✧　テロップ文字のテキスト情報
- ✧　字幕放送のテキスト情報
- ✧　電子番組表のテキスト情報
- ✧　ニュース番組におけるアナウンサーの読み原稿のテキスト情報
- ✧　スポーツ中継番組における試合進行に関するテキスト情報

上記のうち、字幕放送、電子番組表は、本研究当時もデジタル情報として管理されていたが、放送事業者内では、番組制作そのものとは異なる業務であり、システムとしても連携されておらず、実質、メタデータ生成向けに再利用することは困難であった。

最新のファイルベースによるデジタル化されたワークフローは、字幕放送、電子番組表の他、テロップ文字、読み原稿、試合進行情報も加え、全て、同一システム上のデジタル情報として管理され、本研究における区間メタデータと意味メタデータの生成向けに再利用が可能な環境といえる。

これらの情報の中で、特に、テロップ文字、字幕情報や読み原稿のテキスト情報は、豊富なテキスト量と映像内容との一致度から意味メタデータとしてそのまま利用できる可能性が高い。しかしながら、特に、字幕放送のテキスト表示は、リアルタイムより少し遅れて作られるものであり、また、読み原稿に含まれる時間情報はあくまで放送前の予定情報であることから、両情報とも実際の放送番組の内容と時間のズレがあり、区間メタデータへの利用はできない。

本研究のメタデータ生成のターゲットであるニュース番組や野球中継は生放送のことが多いが、上記のように、特に、区間メタデータの生成作業のコスト削減については、最新の番組制作環境においても、有用な情報を事前に生成しておくことは困難である。以上のことから、本研究当時のメタデータ生成の作

業コスト削減に関する課題は、最新のファイルベースの環境においても、産業発展上の意義がある内容といえる。

2.6 まとめ

本章では、映像コンテンツに対するメタデータ生成の作業コストの削減にあたっての各種動向と具体的な課題を整理して述べた。最初に、放送番組のシーン視聴サービスの最新動向として、放送事業者やスポーツ映像の配信事業者の映像配信サービスの実例を紹介した。シーン単位の映像視聴は、従来の家庭内でのパソコンやテレビでの映像視聴に加え、屋外においてスマートホンやタブレットを使って短い時間でも映像が見られるという新しい映像視聴スタイルをもたらすものであると同時に、大きなビジネスチャンスを生むものであることを述べた。

次に、このシーン視聴サービスの実現にあたって、メタデータの必要性、また、その定義を述べた。シーン視聴サービスは、1 つ 1 つのシーン映像を、スポンサー企業の意向を考慮しつつ、1 つの作品として仕上げて提供することが求められるサービスであるが、作品としての一定のクオリティを実現するのに、区間メタデータ、意味メタデータが必要であることを述べた。

シーン視聴サービスは、大きなビジネスチャンスをもたらすものであるが、メタデータは、従来、人手作業で多くの時間をかけて、また、作業者の疲労感も大きい形で作られており、メタデータが増えにくい状況となっていた。この社会的損失の問題を解決するために、メタデータ生成の作業コストを削減する方法が産業活性化のために強く求められていたことを述べた。

更に、メタデータ生成の具体的な作業内容について、従来のアナログベースの番組制作フロー全体における位置づけとともに述べた。作業全体としては、アナログテープの内容をデジタルファイル化する作業もあるが、その作業時間を除いても、メタデータ生成には番組映像の時間長の数倍～10 倍以上かかることがあることを述べた。この問題に対し、メタデータ生成の作業コストの削減を目的とする従来研究として、特に、谷口らの取り組みを紹介した。本研究の課題は、従来研究よりも、より意味的なレベルのメタデータ生成に対する作業コスト削減をターゲットとしたものであることを述べた。

最後に、本研究当時の 2007 年以降、最新の 2018 年時点の番組制作環境としてファイルベースの環境について述べた。ファイルベースは、番組映像に関連

するテロップ文字や字幕テキストが全てデジタル化された制作環境であるが、意味メタデータの生成に対しては、本研究当時よりも、そのまま利用可能なテキスト情報が存在する。しかしながら、区間メタデータの生成作業のコスト削減については、最新の番組制作環境においても、有用な情報を事前に生成しておくことは困難であり、本研究当時のメタデータ生成の作業コスト削減に関する課題は、最新のファイルベースの環境においても、産業発展上の意義がある内容といえることを述べた。

第3章 メタデータ生成の作業コスト削減に向けたアプローチ

3.1 基本的な考え方

本節では、前章で述べた従来技術の動向を踏まえ、本研究でのメタデータ生成の作業コスト削減の実現にあたっての基本的な考え方を述べる。本研究では、映像視聴サービスを提供する事業者の立場に立ち、既に制作済の放送番組をネット配信向けに 2 次利用するケース、また、野球などのスポーツ番組をライブ放送する際に、同時にネット配信するケースを想定し、メタデータ生成の作業コストの削減を検討する。

以降では、本研究のメインテーマである、メタデータ生成の作業コストの削減方法を突き詰めていくにあたり重要と考えられる以下の 4 つの観点で、考え方を整理する。

① 人手作業の必要性

② 作業コストの内容と削減の可能性

③ メタデータ生成の効果的な自動化

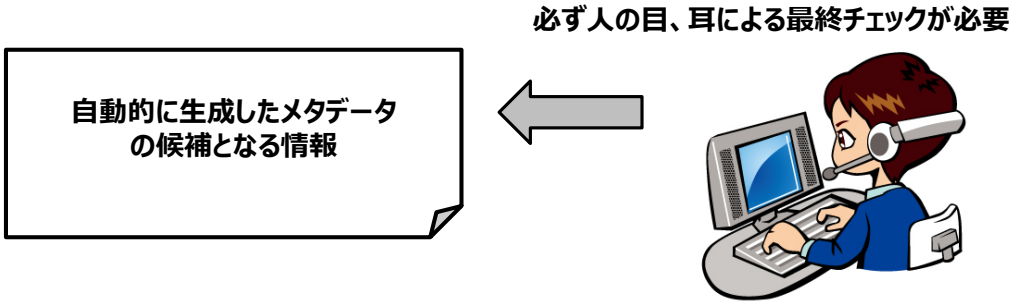
④ ユーザインタフェースのデザイン

3.1.1 人手作業の必要性

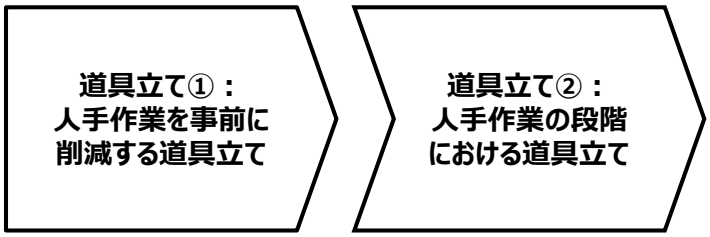
メタデータ生成の作業コストの削減の検討にあたり、まず、重要なこととして、放送番組を対象としたネット配信サービスは、そのサービス要件に対し、スポンサー企業の意向の影響力が強い点、また、エンドユーザに対しては有償でサービスを提供するケースが多い点に注意を払わなければならない。

すなわち、ネット配信する内容に、スポンサー企業のビジネス上、不都合なNG 情報が含まれていないか等、サービス要件が確実に満たされているかを、サービス提供前にチェックする必要がある。このチェックは、映像や音声といったメディア解析技術により、メタデータを自動的に生成する方法がどれだけ高精度になっても、最終的には、メタデータがサービス要件をきちんと満たしているかを、必ず人間の目と耳で最終確認する必要がある（図 3.1.1-1(a)参照）。

従って、本研究におけるメタデータ生成の作業コスト削減にあたっては、人手作業は不可避なものとして、まず、人手でやらなければいけないことを事前に削減する道具立て、次いで、人手作業の段階における道具立ての 2 つの道具立てを実行し、組み合わせることを軸として考えることとする（図 3.1.1-1(b)参照）。



(a) サービス要件に対する最終チェック



(b) メタデータ生成の作業コスト削減のための２つの道具立て

図 3.1.1-1　メタデータ生成に対する人手作業の必要性

3.1.2　作業コストの内容と削減の可能性

メタデータ生成を実行する際に、人手作業は不可避なものであるとして、削減すべきコストの内容を分析し、削減にあたっての考え方を示す。まず、人手作業を行うのは、放送事業者やネット配信事業者において、メタデータ生成の作業を担当する者がこれにあたる。事業者としては、この作業者に支払う人件費、作業用の機器の購入費用、及び、作業スペースがメタデータ生成に必要なコスト全体の構造となる（図 3.1.2-1 の左側参照）。

このうち、機器の購入費用は、例えば、作業用のパソコンと専用のソフトウェア、あるいは、ハードディスクレコーダーといった機器の購入費用となる。パソコンやハードディスクレコーダーは激しい価格競争が起こっており、提供元の企業のビジネス戦略による影響を強く受ける。専用のソフトウェアは、提供元の企業が差異化を図る部分であるが、利用する事業者としては低価格で高機能・高性能なものが望まれるが、この価格も、提供元企業のビジネス戦略の影響を強く受けるものである。また、作業スペースについては、放送事業者、ネット配信事業者等のメタデータ生成を行う事業者の社屋内に存在する物理的空間であり、これは、同事業者の事業計画によるところが大きい要素となる。

これに対し、メタデータ生成の作業者に支払う人件費については、その内訳を分析することで、企業の事業計画等以外の観点から純粋に削減が可能なものである。人件費は、「時間単金」、「作業時間」、「人数」等を掛け合わせたもので構成される。

このうち、「時間単金」は、作業者一人あたりの作業に対する単位時間（例えば１時間）の対価であり、事業者の経営状況や、その時々の市場の相場感などにより、大きく変動する可能性があるものである。一方で、「作業時間」、「作業人数」については、メタデータ生成の作業内容を極力シンプルで分かり易いものとして定義できると、一人当たりの大よその作業時間が見込める可能性がある。そして、一人当たりの作業時間が見込めると、サービス要件に対し、必要な作業人数も事前に見込める可能性が高くなる。

ここで、作業人数を検討する観点として、メタデータ生成の作業全体を複数の小タスクに分割し、各タスクを複数の作業者に割り当て、パイプライン処理で、複数同時実行するような作業スタイルを考える。この作業スタイルの実現のためは、メタデータ生成の作業内容が、明確な小タスクに分割できること、また、複数の作業者の間で、作業時間にあまり個人差が生じないことが成功要因になると考えられる。これに対し、メタデータ生成作業の小タスク化の可能性は、番組内容やサービス要件によって変わりうる点、また、従来から、作業

時間の個人差の程度は明らかになっていない点から、まずは、これらの点を明らかにすることが優先的に検討すべき重要な課題と考えられる。従って、本研究では、複数の作業者による分業スタイルについては、検討のスコープ外として整理し、重要かつ喫緊の課題として、1 人の作業者による作業時間を評価することに主眼を置くこととする。

作業人数の検討にあたり、もう 1 点、コスト、費用、という面では、近年のネット社会におけるクラウドソーシングのように、不特定多数の人数をかけて作業を行うのが最も低コストではないか、という考え方もある。サービス要件によっては、このような方法も有効な場合があると考えられるが、本研究では、社会的信頼度が高い放送番組を扱うことを前提とすることから、信頼のおける作業者により一定以上のクオリティを満たすメタデータを付与することを検討対象とする。すなわち、クラウドソーシング型によるメタデータ生成は、本研究の目的や前提条件にミートしないものと捉え、この考え方についても、本研究の検討スコープ外として整理することとする。

以上のことから、本研究では、コスト構造のうち、作業人数については、複数人ではなく、1 人でメタデータ生成を行うスタイルとして考え、作業コストの削減を検討するにあたっては、作業時間を明らかにすることに主眼を置くこととする。複数人によるメタデータ生成の可能性については、以降の章における具体的な検討を踏まえ、その可能性について議論することとする。

以上を整理すると、本研究では、メタデータ生成のコスト全体構造のうち、人件費の削減に注目し、また、人件費の構成要素のうち、作業単金、作業人数ではなく、作業時間を短縮化することに主眼を置くこととする。すなわち、本研究では、以降、

『メタデータ生成の作業コスト　＝　メタデータ生成にかかる作業時間』

として整理し、この作業時間を、従来の人手作業によるメタデータ生成方式から短縮することを課題として捉える（図 3.1.2-1 右側参照）。

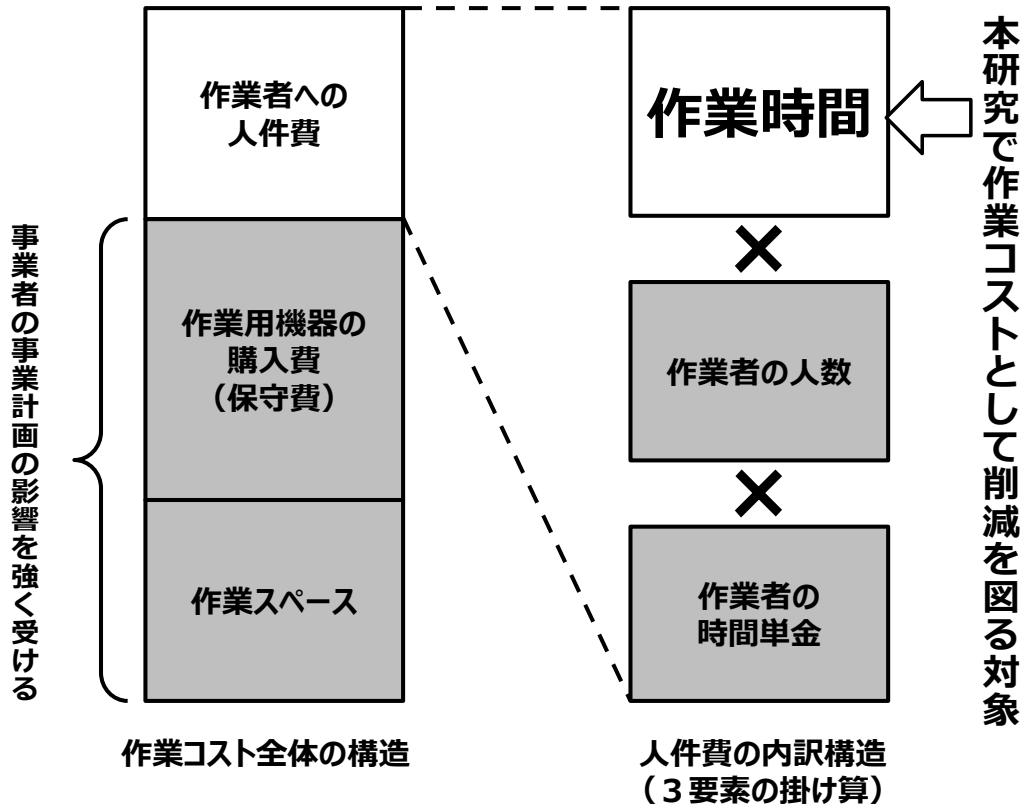


図 3.1.2-1　メタデータ生成の作業コストの全体構造と削減対象

3.1.3　メタデータ生成の効果的な自動化

前項 3.1.2 より、メタデータ生成の作業コストの削減に対し、作業者の作業時間の短縮化を具体的な課題と考えるが、これには、最終的に生成したいメタデータに近い情報を、人手作業の事前になるべく多数得ておくことが重要である。従来技術では、前章で述べたように、映像、音声等の番組コンテンツに含まれるメディア情報を解析するアプローチが提案されており、一定の有効性が認められている[Taniguchi]。本研究でも同様に、人手作業の事前に、メディア解析技術を活用して、メタデータの候補を自動的に得ておくことを考えることとする。

この際、最終的に得たいメタデータに重要な手がかりとなる情報として、映像中のテロップ文字情報に着目することとする。区間メタデータ、意味メタデータのうち、特に、意味メタデータのほうは、テキスト情報であることから、メディア解析をする場合に、映像信号の中のテロップ文字や音声がテキスト化できると非常に有用である。

本研究では、図 3.1.3-1 に示すように、テロップ文字が持つ意味メタデータの手がかりとしての可能性に加え、特に、従来技術では見られない区間メタデータとしてのテロップ文字の有用性に着目する。テロップ文字は、例えば、ニュース番組中の各ニューストピックの冒頭に表示されるニュースタイトル、また、スポーツ中継での盛り上がりのシーンのタイミングで表示される選手名や得点数字等、番組中のシーンの意味的な区切りのタイミングで用いられることが多い。

具体的な方法は次章で述べるが、本研究では、区間メタデータ、意味メタデータ、両方の手がかりとなる重要な情報を効果的に得る手段として、テロップ文字を映像信号から検出し、認識する方法を重要視して考えることとする。また、意味メタデータは、前章で述べたが、ニュース番組の場合、アナウンサーの読み原稿などの既存テキスト情報も有用であることから、これらを再利用することも検討のスコープに入れて考えることとする。

図 3.1.3-1　テロップ文字のメタデータとしての有用性

3.1.4　ユーザインタフェースのデザイン

メタデータ生成のプロセスの中で不可避な人手作業向けのユーザインタフェースのデザインに関する考え方を述べる。作業時間の短縮化を図るべく、従来の人手作業の方法を分析し、特に時間がかかる作業を抽出し、デザインを検討する。

メタデータ生成の全てを人手作業で行う従来方法は、デジタルコンテンツとして存在する映像コンテンツをハードディスクレコーダーやパソコン上のデジタルファイルとして取扱い、各デバイス上の映像再生プレーヤー、及び、早送り、巻き戻し、コマ送り、コマ戻し等の再生制御機能を駆使しつつ、映像内容を目視や耳で確認しながら実行するものである。

この作業は、メタデータを生成するシーンを見逃さないように、映像、音声を注意深く見聞きしなければならない。作業時間の短縮化のために早送り再生をしても、メタデータ生成対象のシーンがいつ表示されるか分からないことから、神経を集中させ、映像内容を注意深く確認し続ける必要がある。このように、全て人手作業で行う場合は、映像内容を最初から最後まで目視しなければならず、映像を再生している時間の長さが作業時間に大きく影響する。また、作業者の疲労感も大きいものである（図 3.1.4-1 参照)。

図 3.1.4-1　人手作業によるメタデータ生成の問題点

これに対し、テロップ文字を始めとする映像特徴をメタデータの手がかりとして、人手作業の事前に抽出する提案方式におけるユーザインタフェースとしては、極力、映像の再生時間が少なく済むようなデザインを重視する。すなわち、映像を再生することなく、なるべく、映像内容、音声内容の全体像を俯瞰できるようなデザインが望ましい（図 3.1.4-2 参照）。

具体的には、区間メタデータの生成に対しては、人手作業の事前に抽出した区間メタデータの候補タイミングになりそうな瞬間の静止画像をなるべく多数、時系列に沿って一覧表示できるようにする。また、音声のレベルについても、時系列に沿って俯瞰できると、再生せずに、音声の有無が確認できる。

意味メタデータの生成に対しては、シーンのタイトル、概要、キーワードといったテキスト情報に関し、従来は、映像、音声を見聴きしながら、一から書き起こしていた。これを短時間化するには、一から書き起こす必要なく、文字認識や音声認識の結果が最初から確認できるデザインが望ましい。また、文字認識や音声認識の結果のテキスト情報は、ニュース番組の場合、膨大なテキスト量となる。この膨大なテキスト情報に対し、該当箇所を目視で探す必要なく、すぐに該当箇所の確認、必要に応じた修正作業ができる状態が必要と考える。

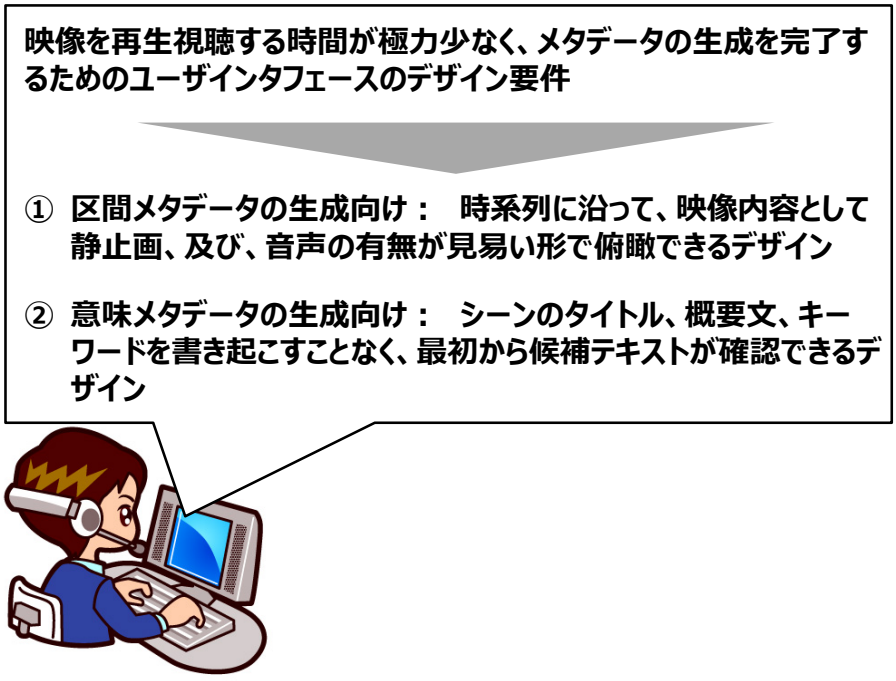


図 3.1.4-2　ユーザインタフェースのデザイン要件

3.2　提案方式のアプローチ

前節で述べた下記の各観点①～④の考え方を踏まえた、本研究における提案方式のアプローチについて述べる。

① 人手作業の必要性
② 作業コストの内容と削減の可能性
③ メタデータ生成の効果的な自動化
④ ユーザインタフェースのデザイン

まず、①については、放送番組のネット配信に関するサービス要件より、人手作業によるメタデータ生成の最終チェックは必須である点を注意する必要がある。具体的には、図 3.1.1-1(b)に示したような、大きくは、「人手作業を事前に削減するための道具立て」と「人手作業を削減する道具立て」の 2 つの道具立てを作り出すことを考える。この大きな枠組みに、②、③、④の考え方を反映したメタデータ生成の作業コスト削減方式、及び、評価方法を検討する。

②の考え方の反映については、3.1.2 項で述べたように、メタデータ生成の作業コストの削減に対しての具体的な方法として、作業者の「作業時間」を短縮化することを考える。すなわち、提案方式の評価方法として、この作業時間を計測し、結果を考察することとする。次に、③の考え方の反映として、前述の「人手作業を事前に削減するための道具立て」に、3.1.3 項で述べたように、番組コンテンツの映像、音声を解析し、メタデータの手がかりを得る。この際、特に、メタデータの手がかりとして有用な映像中のテロップ文字を認識することを特徴としてアプローチを取る。そして、④の考え方との反映として、「人手作業を削減する道具立て」に、3.1.4 項の図 3.1.4-2 で示した、ユーザインタフェースのデザイン要件を満たす仕組みを導入する。

以上のことをまとめた内容を図 3.2-1 に示す。本研究の提案方式を 2 ステップで構成することとし、第 1 ステップで、映像、音声、言語処理といったメディア解析処理により区間メタデータ、意味メタデータの手がかりとなる情報を生成する。その際、特に、テロップ文字の解析により最終的に映像視聴サービスに利用できるメタデータとほぼ同様の内容を自動的に得ることを考える。第 2 ステップでは、メタデータの手がかりとなる情報を、図 3.1.4-2 で示すユーザインタフェース要件を満たす仕組みで表示し、最小限の人手作業により、メタデータを完成させる。

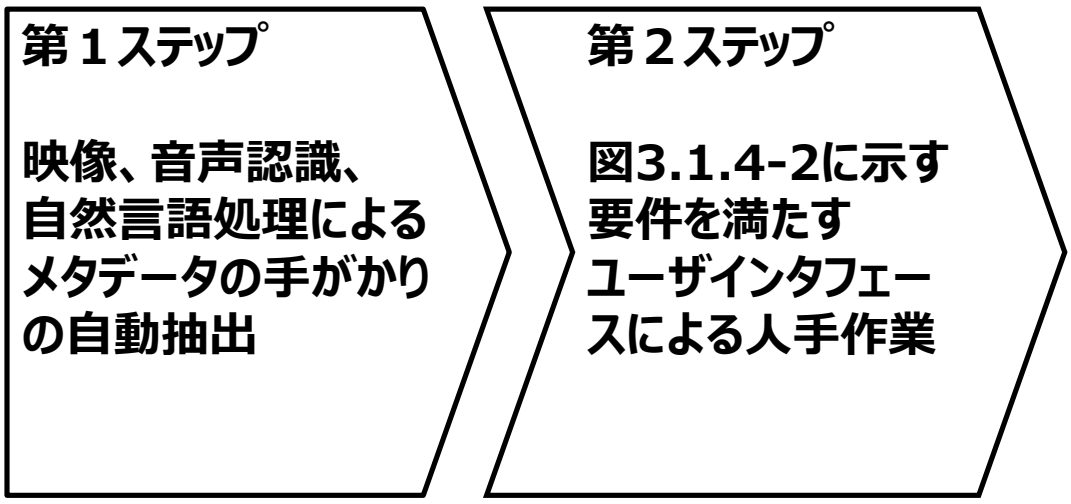


図 3.2-1　提案方式の 2 ステップ

3.3　具体的な個別検討テーマ

前節 3.2 で述べた提案方式の 2 ステップを具現化していくにあたり、具体的な個別検討テーマを設定する。メタデータ生成作業のコスト削減という課題に対し、従来技術では解決できない新たな産業インパクトをもたらす可能性が高いと考えられる幾つかのテーマを設定する。

本研究の主目的はメタデータ生成の作業コスト削減であるため、ストレートには､番組コンテンツに対するステップ 1､ステップ 2 の具現化方法を策定して、その方法が、実際にどの程度人手作業の時間を短縮するかを評価するテーマが考えられる。これに対し、メタデータの生成対象の番組コンテンツの種類として、制作済番組とライブ番組の 2 種類存在し、サービス要件にもよるが、この 2 種類の番組種類の間で、人手作業の内容が大きく異なることが考えられる。

すなわち、制作済番組に対するメタデータ生成は、コンピュータによる自動処理で実現可能なステップ 1 を夜間のバッチ処理で実行し､ステップ 2 として、その結果を作業者が最終的なメタデータとして完成させる作業モデルがとれる。ステップ 1 の処理は多少時間がかかる場合でも、夜間に実行することで、作業者の待機時間等の削減が見込める。

一方、ライブ番組に対するメタデータ生成は、番組の配信中に、番組の進行に遅れることなく、同時にメタデータ生成を実行し、配信をする必要がある。

図 3.3-1 に、野球中継を例として、ライブ番組のメタデータを活用したダイジェスト視聴サービスの画面例を示す。画面には、番組の進行に合わせて、ダイジェスト視聴用のシーンに関するメタデータが時々刻々増えて表示されていくイメージである（図 3.3-1 の画面の左下)。番組視聴者は見たいシーンに関する情報を選択すると、そのシーンのリプレイ映像が視聴できる、いわゆるタイムシフト視聴が番組進行中に実行できるものである。このようなライブ番組へのメタデータ生成の実現には、夜間バッチ処理という概念は存在せず、作業者は、番組放送中にリアルタイムに区間メタデータと意味メタデータを生成しなければならない。

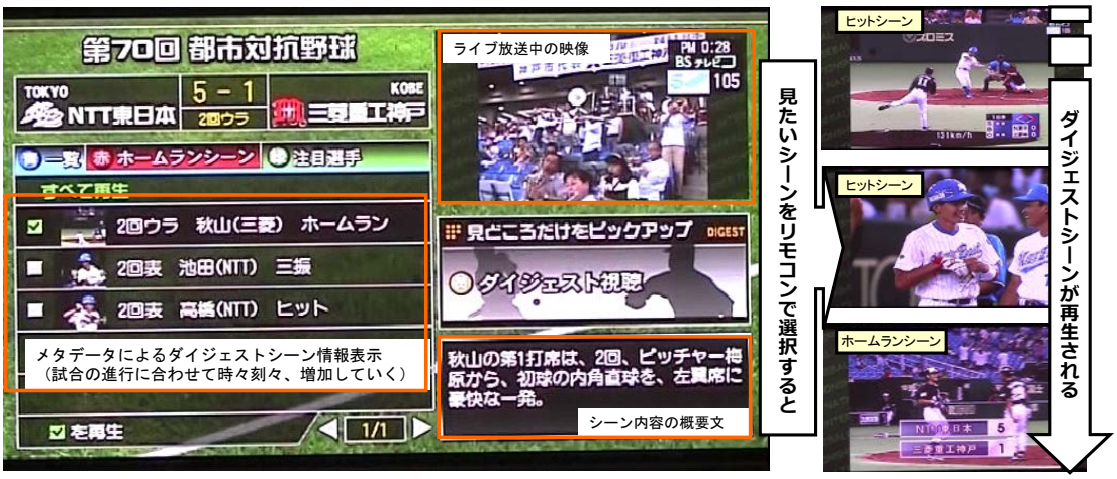


図 3.3-1　ライブ野球中継のダイジェスト視聴サービスの画面例

以上のことから、本研究における個別検討テーマとして、制作番組向けのメタデータ生成、及び、ライブ番組向けのメタデータ生成の 2 つを設定する。

加えて、メタデータ生成の作業コスト削減という主目的に対しては、手段という位置づけになるが、テロップ文字認識の新しい方法について、特に、区間メタデータの手がかりとしてテロップ文字の大きさや表示位置に着目する方法は従来存在しないものであることから､これも個別テーマの１つとして設定し、検討を進める。

以降の章では、前述の 3 つの検討テーマに沿って、以下の構成とする。

- テロップ文字の認識によるメタデータの自動生成の方式提案と実験、考察を述べる。従来技術には存在しない、テロップ文字の大きさや表示位置といった視覚特徴を用いた区間メタデータの自動生成が特徴である。
  ⇒　第 4 章で述べる。
- テロップ認識も活用した上での、制作済番組向けのメタデータ生成の作業コスト削減について述べる。メタデータ生成のユーザインタフェースシステム「SceneCabinet / NBS」を提案し、同システムを用いたメタデータ生成と人手作業によるメタデータ生成の作業時間の比較実験の結果と考察を述べる。　⇒　第 5 章で述べる。
- ライブ番組向けのメタデータ生成の作業コスト削減についての方式提案、実験、考察を述べる。ライブ番組向けのメタデータ生成のユーザインタフェースシステム「SceneCabiet / Live!」を提案し、同システムを用いたライブ番組向けのメタデータ生成と人手作業によるメタデータ生成の作業時間の比較実験の結果と考察を述べる。　⇒　第 6 章で述べる。

3.4　まとめ

本章では、本研究における番組映像のシーン視聴サービスに必要なメタデータ生成の作業コスト削減の実現にあたっての基本的な考え方とアプローチを述べた。基本的な考え方としては、サービス要件や従来のメタデータ生成の作業コストに関する問題点から、「人手作業の必要性」「作業コストの内容と削減の可能性」「メタデータ生成の効果的な自動化」「ユーザインタフェースのデザイン」の各観点から述べた。

人手作業の必要性の観点からは、放送番組のネット配信サービスの要件から、人手作業のコストを削減することが本研究の課題ではあるが、本研究で提案する方法としても、最終段階では、人手によるメタデータ内容の確認は必要であることとした。すなわち、この最終段階の人手作業をいかに削減するか、という問題として、以降の検討を行うこととした。

作業コストの内容と削減の可能性の観点では、最初に、コスト全体の内容として、作業者の人件費、作業用機器の購入費、作業スペースの確保を挙げた。これらに対し、メタデータ生成を行う事業者の事業計画に依存する要素は、本

研究におけるコスト削減の対象外とし、結果として、本研究におけるメタデータ生成の作業コストとは人手作業の作業時間を短縮することとして整理した。

メタデータ生成の効果的な自動化の観点では、シーン視聴サービスに利用可能な最終的な区間メタデータ、意味メタデータに近い情報が人手をかけずに得られる方法を実現する必要がある。本研究では、映像中のテロップ文字の持つ、映像内容との意味的、時間的な関係が強いという特徴に着目し、画像処理によりテロップ文字の情報から区間メタデータ、意味メタデータの手がかりを作り出すこととした。

ユーザインタフェースの観点では、メタデータ生成の最終段階で人手作業が短時間で実行できることを考慮したデザインにする必要がある。従来の人手作業で時間がかかっている、映像の再生、早送り、巻き戻し等の操作を極力行わずに済むデザインとして、映像中の静止画や関連のテキスト、あるいは、音声情報の有無を可視化することが重要であることを述べた。

以上の整理事項に基づいて、本研究におけるメタデータ生成の作業コストの削減に向けたアプローチとして、以下の 2 ステップで構成する方法を提案することを述べた。ステップ 1 は、映像中のテロップ文字の抽出、認識を中心とし、映像、音声認識、自然言語処理といったメディア解析技術を活用し、メタデータの手がかりとなる情報を得るステップ。ステップ 2 は、ステップ 1 の実行から結果の表示、及び、その内容を確認して、必要に応じ修正し、最終的なく関メタデータ、意味メタデータを出力する作業者向けのユーザインタフェースを活用するステップであることを述べた。

2 ステップから構成される提案手法を具体化していくにあたり、まず、テロップ文字の抽出、認識に関する手法の提案と評価も個別検討課題とした。そして、映像のシーン視聴サービスが、制作済の番組を対象とする場合とライブ番組を対象とする場合とでメタデータ生成の要件が異なることから、それぞれを個別の検討課題とすること、以上を本研究の基本的考え方とアプローチとして述べた。

第4章 テロップ文字認識によるメタデータ自動生成

4.1 概要

本章では、メタデータの手がかりとなる情報を自動的に生成するアプローチとして、映像中のテロップ文字を認識する方式を提案し、その実験結果、及び、考察を述べる。提案方式は、映像中のテロップ文字の元となっているテキスト情報の他、テロップ文字の大きさや表示位置など、テロップ文字の持つ様々な情報をメタデータ生成に活用する方式となっている。特に、テロップ文字の大きさや画像中の表示位置と、番組映像の意味内容の相関に着目し、区間メタデータを生成する方式は、従来存在しない新たな試みとなる。

以降、4.2 節では、テロップ文字認識に関連する従来技術、また、テロップ文字以外の映像特徴として、カット点、カメラワーク、音楽等に着目したメタデータの手がかりとなる情報の抽出に関する技術を整理して述べる。これらの従来の技術が、本研究で掲げる課題を十分には解決できないという問題点を述べる。

4.3 節では、本研究におけるテロップ文字認識に関する提案方式として、映像中からテロップ文字が表示されている画像を検出する方式、この画像からテロップ文字の画素だけを抽出する方式、その後、テロップ文字の大きさ、表示位置を検出し、1 文字ずつ切り出し、文字認識処理を行う一連の方式を順に述べ、合わせて、実験結果、考察を述べる。

4.4 節では、テロップ文字認識処理の中で得られるテロップ文字の大きさ、表示位置といった視覚的な特徴を活用し、また、これに加え、映像中のカット点や人物動作等、他の映像特徴を組み合わせた、従来技術には存在しない区間メタデータの自動生成方式を提案する。

4.5 節では、提案方式の応用例として、テレビ番組全チャンネル同時録画検索システム「Telop on demand」について、その実装方法を中心に述べ、4.6 節で本章のまとめを述べる。

本章の内容は、筆者らの公表論文[C1],[C2],[C3]の内容に基づくものである。

4.2 従来技術

4.2.1 テロップ文字認識

テレビ番組中のテロップ文字は、番組内容の意味的な情報を提示する役割を担う。このため、テロップ文字の元のテキスト情報を映像全体のシーン検索のインデクスとして抽出する技術が、特に、1997 年～1999 年に、国内外の論文として多数発表されている[Smith],[Gargi],[Ariki],[Lienhart],[Shim],[Hori],[Mori]。

いずれの従来技術も、映像信号に対し、テロップの文字らしさの特徴量を定義し、条件を満たすものをテロップ文字として抽出し、テロップ文字が表示された時間やテロップ文字の認識結果のテキスト情報をメタデータの手がかりとして利用するものであった。背景映像に対し、高い輝度コントラストで表示され、また、一定の解像度以上のテロップ文字に対して、表示タイミングやテキスト情報を正しく取得できるレベルであった。

一方で、特に、区間メタデータの手がかり情報の抽出という観点からは、従来技術には大きな問題点がある。従来技術で得られるテロップ文字の出現タイミングの情報は、映像全体の中のテロップ文字が表示されている全てのシーンが対象である。これに対し、本研究における想定ターゲットサービスであるシーン視聴サービス向けの区間メタデータは、ニュース番組の各ニューストピックの冒頭シーン等、意味のあるシーン区間のみを取り出し、その区間の先頭の時間情報である。

映像全体の中から、意味のあるシーンの開始タイミングだけを選択的に抽出する技術の検討、また、意味メタデータも含め、メタデータ生成の作業コストの削減を課題としたテーマは､本研究当時の学会論文誌や著名な国際会議（ACM Multimedia, IEEE Multimedia）においても発表が無い。すなわち、従来技術は、本研究が課題とするメタデータ生成の作業コストの低減に対し、特に、区間メタデータの生成作業に関しては、コスト削減の効果が大きくは期待できず、大きな問題と考えられる。

4.2.2　テロップ文字認識以外の従来技術

本項では、メタデータ生成の手がかりとなる情報を自動生成する技術に関し、テロップ文字認識以外のアプローチについて、本研究の周辺動向として述べる。1995 年頃から、大量の映像データに対し、映像のシーン内容に踏み込んだ検索、いわゆるシーン検索の機能を備える映像アーカイブスの構築が放送事業者、博物館、自治体など、膨大な映像データを扱っている事業者、組織等で注目を集め始めた。

映像アーカイブスにおけるシーン検索用のインデクスを自動的に得る方法として、テロップ文字以外にも、カット点、音声情報、被写体の動き等を映像信号中から取り出す方法が多数提案されていた。

カット点は、シーンが切り替わるタイミングであり、それ単体の情報では映像内容の意味までは判別困難であるが、区間メタデータを生成する際に、映像全体のおおよその流れ、ストーリー展開を把握する上では有用な情報である。カット点検出は、非常に多くの研究がなされており、映像信号の特徴として輝度や色のヒストグラムを活用するものが多い。一般的なシーン切り替えの他、切り替え時にフェード効果が施された方法、あるいは、カメラのフラッシュが多数たかれても誤検出しないような工夫が施された方法等、多数提案されている[Taniguchi],[Kawai],[Kumano],[Suzuki],[Miura],[Takimoto]。技術レベルとしては、正しくカット点検出できる精度が、ほぼ 100%に近いくらいところまで到達しており、成熟している技術と捉えられる。

番組コンテンツ内の音声についても、これを認識してテキストデータにすることで意味メタデータの有力な手がかりとなることもあり、従来から多数の取り組みがある[Andou],[Bessho],[Kobayashi]。これらは、ニュースのアナウンサー音声を対象とした取組であり、ニューススタジオでのアナウンサー音声など、背景ノイズ音が無い場合に、文章としては 80%以上、単語の正答率でみると 90%以上の正しい認識精度がでるレベルである。しかしながら、ニュース以外の番組になると、一気に精度が低下してしまうのが現状である。例えば、スポーツ番組のアナウンサー音声は、野球のホームラン等、盛り上がるシーンでは、絶叫に近い音声が、ノイズ音の上に重畳され、機械学習をベースに行う方式では、正解クラスタが作りにくいことから、まだ技術課題が多いレベルといえる。

他に、複数のメディア解析技術を組み合わせ、シーン検索のインデクスを作り出す方法も提案されている。[Miura]はカット検出と人物認識を組み合わせて、調理番組の構造化を試みている。[Fujimoto]は、テレビショッピング等の商品紹介番組を対象に、出演者の音声、及び、商品を説明するテロップ文字を認識し

た結果を組み合わせている。音声認識の辞書データに対し、テロップ文字の認識結果を反映することで音声認識の精度向上を図るアプローチである。[Kobayashi]もニュース番組の音声認識にあたり、辞書データにウェブ上のニュース記事のテキストを反映することで音声認識の精度向上を図っている。[Yoshida]は、歌番組を対象に、映像中の顔認識、音響特徴抽出を組み合わせ、番組のシーン分割を試みている。[Mikami]は、野球番組に特化したメタデータ生成として、投球動作の検出と突発音の検出（例：キャッチャーミットにボールが収まる際の音、バットにボールが当たった際の音）を組み合わせ打席シーンの抽出を試みている。

以上が、テロップ文字以外の映像、音声の解析技術によるメタデータの手がかり情報の自動生成に関する動向である。従来技術は、各メディア解析の手法そのものの高精度化に主眼が置かれ、メディア解析の精度そのものが評価対象になっている。すなわち、本研究が課題としている、シーン単位の映像配信サービスに利用可能なメタデータを生成する作業まで含めたコスト評価を行っている取組は存在しない。

4.3　テロップ文字認識方式

4.3.1　概要

本節では、メタデータ生成の作業コスト削減にあたり、有用な情報になると考えられる、テロップ文字の認識の方法を提案し、実験結果と考察を述べる。映像信号の持つ輝度や彩度等の特徴量に対し、テロップ文字としての条件を策定、適用し、テロップ文字が表示されるタイミング、テロップ文字の大きさ、表示位置を抽出し、文字認識処理を行い、区間メタデータ、意味メタデータの候補を自動生成する。提案方式を実際のニュース番組に対して適用した実験結果を述べると共に、テロップ文字をシーン単位の映像視聴サービス向けのメタデータに活用することの有効性を明らかにする。

図 4.3.1-1 に方式の全体像を示す。映像信号データを入力とし、区間メタデータ、意味メタデータの候補を出力とする方式である。方式の全体的な流れとしては、映像信号データに対し、最初に、テロップ文字が表示されるタイミングを検出し（①テロップ画像検出）、次に、テロップ文字が表示されるフレーム画像中の文字領域を抽出し（②テロップ領域抽出）、次に、テロップ領域から一つ一つの文字列、及び、文字サイズや表示位置を抽出し（③テロップ文字列抽出）、

最後に、抽出した文字領域をテキストコード化する文字認識を行う（④テロップ文字認識）。

以降では、4.3.2 項で、テロップ画像検出、4.3.3 項で、テロップ領域抽出、4.3.4 項で、テロップ文字列抽出の方式提案と実験結果、考察を述べる。なお、図 4.3.1-1 の④テロップ文字認識については、意味メタデータの生成に必要であるが、本研究では本処理はスコープ外とし、[Mori]の方式を利用するため、以降では、テロップ文字列抽出までを述べる。

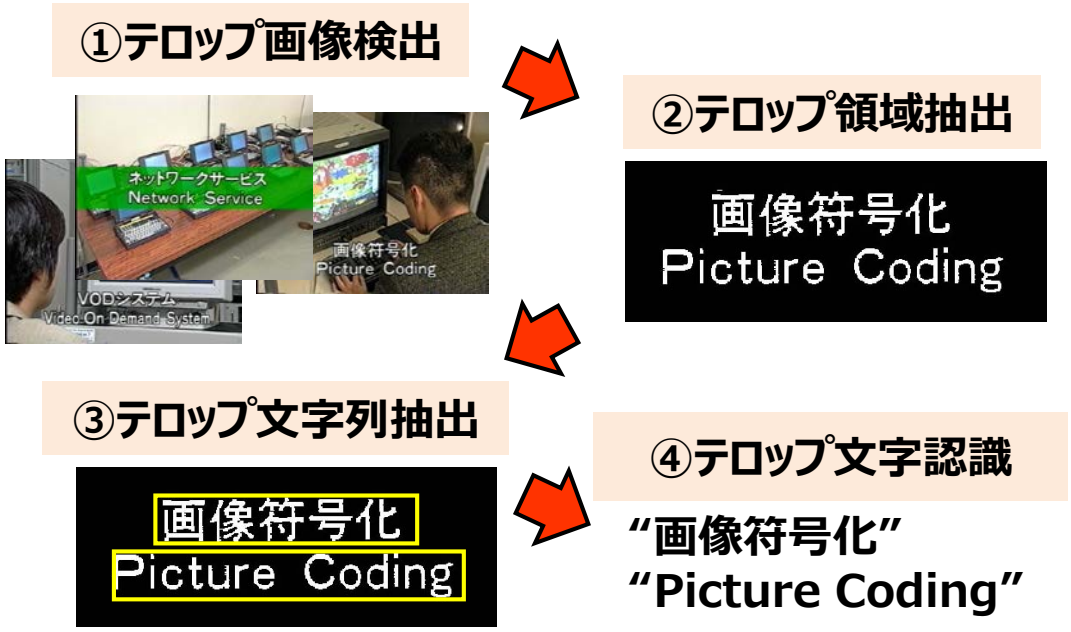


図 4.3.1-1　テロップ文字の検出・認識の処理フロー

4.3.2　テロップ画像検出

4.3.2.1　概要

本項では、メタデータの手がかりとなる映像特徴として、テロップ文字が表示されている画像（以降、テロップ画像と呼ぶ）の自動検出の方法について本研究での提案内容を述べる。映像信号中の輝度パターン情報に対し、テロップ文字らしさの特徴量「エッジペア」を定義し、従来技術では困難であった、輝

度コントラストが高い複雑な背景とテロップ文字を正しく区別し、高精度にテロップ文字画像を検出する方法として提案する。実験では従来技術との比較を行い、提案方式の有効性を示す。

4.3.2.2 従来技術

映像中からテロップ画像を自動検出する方法として、従来、画像全体の中の局所領域における輝度ヒストグラムや輝度エッジ画素数の時間変化に注目した手法が提案されている[Nemoto],[Nakajima]。これらは、テロップ文字列としての空間配置の特徴が考慮されておらず、十分な精度が得られなかった。

一方、画像から検出された輝度エッジの局所的な分布形状と時間的な継続性に注目した手法[Smith] では、背景の輝度エッジとテロップ文字の輝度エッジが分離できない場合に正しく検出されない。

このほか、圧縮符合化された映像データを対象とし、テロップ文字表示中の符号化データの特徴に着目した手法が提案されている[Satoh], [Takano]。これらはデコードせずにテロップ文字を検出できる利点があるが、符号化データだけでは十分な検出精度が得られない。

このように、従来の手法ではテロップが表示されるフレーム画像を検出する精度が不十分であるという問題があった。

4.3.2.3 アプローチ

映像中のテロップ画像を従来技術よりも背景画像を誤検出しないための手法を提案する。提案手法では，テロップ文字と背景の輝度エッジのパターンの違いに着目することにより、テロップ画像を検出する。

本研究では、画像中の輝度分布に対し、テロップ文字周辺でみられ、かつ背景物体からはみられない特徴点として「エッジペア」を提案する。従来手法でも、特徴点として輝度エッジを用いているものが多いが、輝度エッジはテロップ文字以外の背景物体からも検出されるため、テロップと背景物体の区別が困難な場合も多い。

図 4.3.2.3-1 に、「エッジペア」の定義とテロップ画像の判断基準を示す。テ

ロップ文字の輪郭部に存在する輝度エッジは、横方向及び縦方向のライン上で、逆勾配の関係を持つ隣接する 2 エッジを「エッジペア」と定義する。このエッジペアが空間的に密集し、一定時間、同じ位置に存在し続けたらテロップ文字が表示されているものとして検出する。

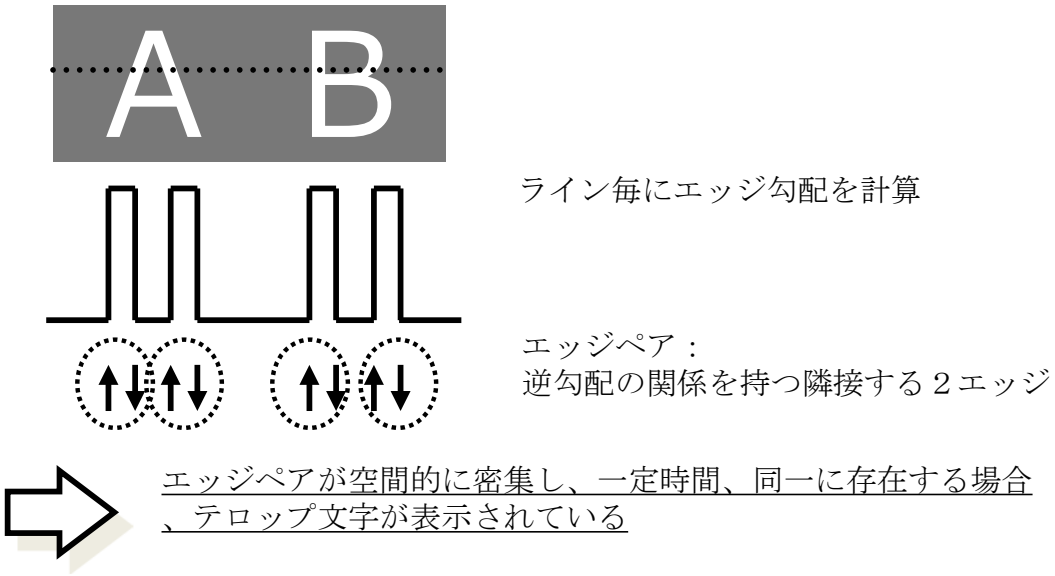


図 4.3.2.3-1 エッジペアによるテロップ文字表示画像の検出

4.3.2.4　提案方式

エッジペアを活用したテロップ画像の検出方法の具体的な方法を提案する。

最初に、ある時点でのフレーム画像 Fi において輝度エッジを検出した後、輝度エッジ画像からエッジペアを検出する。図 4.3.2.4-1、及び、図 4.3.2.4-2 にエッジペアを検出した例を示す。

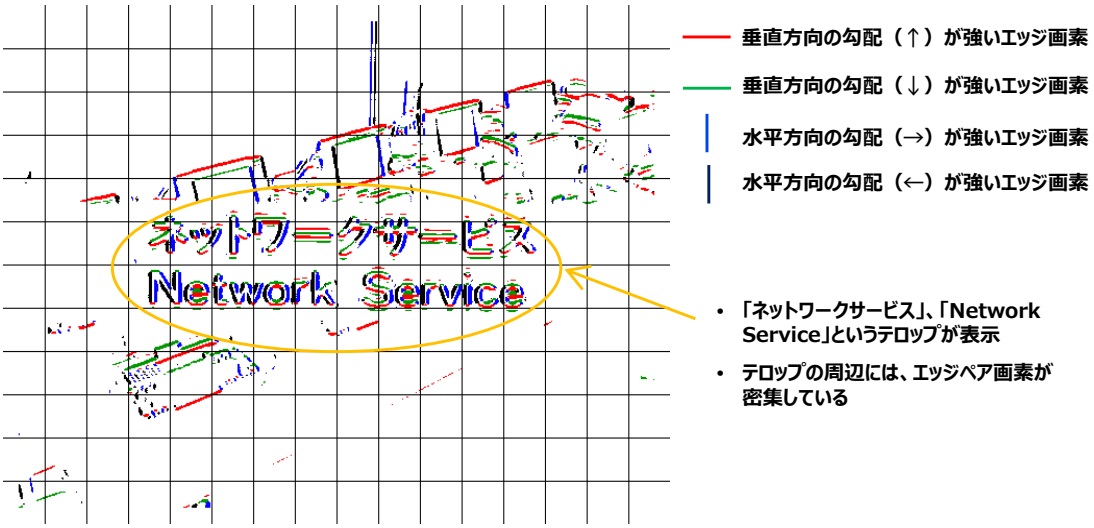


図 4.3.2.4-1：　エッジ画像の例（テロップが表示されている画像の場合）

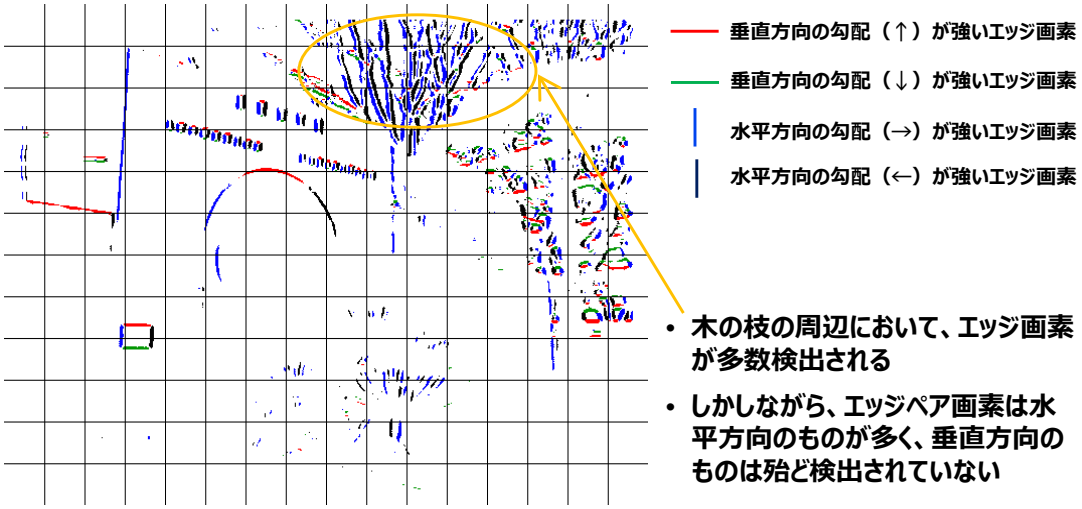


図 4.3.2.4-2：　エッジ画像の例（テロップが表示されていない画像の場合）

次に、エッジペアを検出した画像全体を、想定テロップ文字サイズと同じくらいの大きさの M×N 個、面積 s の部分ブロックに分割する。そして、ブロック毎に垂直方向のエッジペア数 ρ_vと水平方向のエッジペア数 ρ_hを数える。部分ブロック内のエッジペア数が一定以上存在し、かつ、水平方向と垂直方向のエッジペアの存在比が一定以上の場合、その部分ブロック内にテロップ文字の一部が存在する可能性が高いといえる。これを以下に示す、エッジペア密集度条件により評価する。

$$\rho_0 \leq \frac{\rho_h + \rho_v}{s}$$ かつ、

$$r_0 \leq \frac{\rho_h}{\rho_v} \leq \frac{1}{r_0}$$ ・・・エッジペア密集度条件(1)

ここで、 ρ_0及び r_0 $(0 < r_0 < 1)$ は閾値である。エッジペア密集度条件を満たす部分ブロックを、テロップ候補ブロックと呼ぶこととする。次に、検出されたテロップ候補ブロックの画像全体における大域的な分布を評価する。これは、画像全体を部分ブロックで分割した際の、部分ブロック行 m、または、部分ブロック列 n の中に、テロップ候補ブロックが一定個数以上存在する場合、そこに、テロップ文字列が存在している可能性が高いと考える。これを以下に示す文字列表示条件を用いて評価する。

$$A_m^h = \sum_{k=1}^{N} b_{m,k} \geq A_0$$ または、

$$A_n^v = \sum_{k=1}^{M} b_{k,n} \geq A_0$$ ・・・文字列表示条件(2)

ここで、$b_{m,n}$ はブロック(m, n) のエッジペア密集度条件(1) の判定結果（=1：テロップ候補ブロック、=0：テロップなし）であり、A_m^h 及びA_n^v は、部分ブロック行 m、 及び部分ブロック列 n に含まれるテロップ候補ブロックの個数、A_0 はその個数に関する閾値である。いずれかの部分ブロック行、または部分ブロック列が条件（2）を満たす場合、Fi にテロップ文字が表示されている可能性

があると判定する。以上は、1 枚のフレーム画像内に閉じた処理であるが、この処理を連続するフレーム画像で行い、文字列表示条件が時間的にどのように変化をするかをみて、最終的にテロップ文字画像かどうかを判断する。

文字列表示条件の時間的な変化の評価にあたっては、部分ブロック毎のエッジペア点の継続性の評価を行う。フレーム画像 Fi のテロップ候補ブロック(m, n) 内の全てのエッジペア点について、後続のフレーム画像 Fj 上の同じ位置に、同じ勾配方向をもつエッジペア点を数え、密集度条件(1) を満たす場合、その部分ブロックにはテロップ文字らしい表示が継続していると判定する。この評価を、更に後続のフレーム画像との間で繰り返す。文字列表示条件(2)を満たす部分ブロックで継続性の条件を満たさなくなった場合、そのフレーム画像のタイミングでテロップ文字が消滅したと判断する。

初めて文字列表示条件(2)を満たしたタイミングをテロップ文字の開始時間、消失したタイミングを終了時間として検出する。

4.3.2.5 実験結果と考察

テロップ画像の検出に関する提案方式の精度評価は、8 種類のニュース番組の映像、約 10 時間分を用いて行った。映像の解像度は 640 画素×480 画素である。提案方式の各種パラメータは予備実験により経験的に決定した。部分ブロックのサイズは 40×40 画素（ブロック数 12×16）、$\rho_0 = 0.015$、$r_0 = 0.1$、$A_0 = D_0 = 2$ とした。時間的な継続性の評価の際の後続のフレーム画像との時間間隔は約 0.4 秒とした。また、テロップ文字の最短の表示時間を 2 秒と仮定した。

実験結果を表 4.3.2.5-1 に示す。正検出率 95%であり、ニュースの冒頭のヘッドライン等の重要なテロップ文字ほど安定して検出できる傾向があった。検出漏れは画面の隅に小さく表示される注釈（日付等）など、比較的重要度が低いテロップ文字場合が多かった。部分ブロックを 1 種類固定としたため、多様な文字サイズへの対応が今後の課題となる。また、エッジペアの適用により、テロップから検出されるエッジの多くを残した状態で背景からのエッジを削減することができ、背景画像中の樹木等の自然物、人混みの遠景等の誤検出を抑えることができた。

表 4.3.2.5-1　テロップ画像の検出結果

		値
テロップ画像の合計の枚数		１，３８３
結果	正検出された枚数	１，３１４　（９５%）
	検出漏れした枚数	６９
	誤検出した枚数	１１１

4.3.3　テロップ領域抽出

4.3.3.1　概要

前項 4.3.2 で述べた、テロップ画像は映像中から検出したタイミングが区間メタデータの手がかりとなる情報である。また、テロップ文字は文字認識をすることで、意味メタデータの手がかりにもなる。本項では、このような区間メタデータ、意味メタデータの更なる手がかりを得るための画像処理の課題として、テロップ画像から画像中のテロップ文字領域の自動抽出の方法を提案し、実験結果と考察を述べる。

本研究における、テロップ文字領域の自動抽出とは、前項で述べた手法で検出した、テロップが表示されているカラー画像を入力画像とし、画像中のテロップの画素だけを抽出し、テロップ画素を白色、それ以外を黒色の二値画像として出力する処理である。我々が提案する方法は、映像中の各走査線に沿って生じる各画素の輝度や色の滲み現象に対しても、高精度にテロップ画素を抽出することを指向したものである。

4.3.3.2　従来技術

文字が表示されるカラー画像から文字部分だけを抽出する画像処理技術は従来から、景観画像中の看板文字や標識文字、あるいは、書籍の表紙、CD カバー上のデザイン文字の抽出など多数の研究がなされている[Hori]。

図 4.3.3.2-1 に示すように、従来の技術はいずれも、文字色が文字内部で一様な輝度値をとることが前提とされており、画像全体の中でも、文字周辺の局所的な範囲における輝度ヒストグラムの閾値処理を用いた二値化により文字画素を抽出していた。

本研究で処理対象とする放送番組の映像は、各走査線に沿って各画素の輝度や色が滲む性質をもち、テロップについては背景との左右の境界部で背景部の輝度が文字内部に滲み本来の輝度が劣化する場合がある。図 4.3.3.2-1 に示すが、白い「t」というテロップ内部に背景の黒色が滲みこみ、「t」の縦ストロークについては、文字幅が狭いために全体が劣化して輝度が低下する。これにより、文字内部で、横ストロークと縦ストロークで輝度の値が大きく異なる。このような場合、従来の二値化方法によると、縦ストロークが正しく文字として抽出されず、文字が途切れてしまう結果となる。

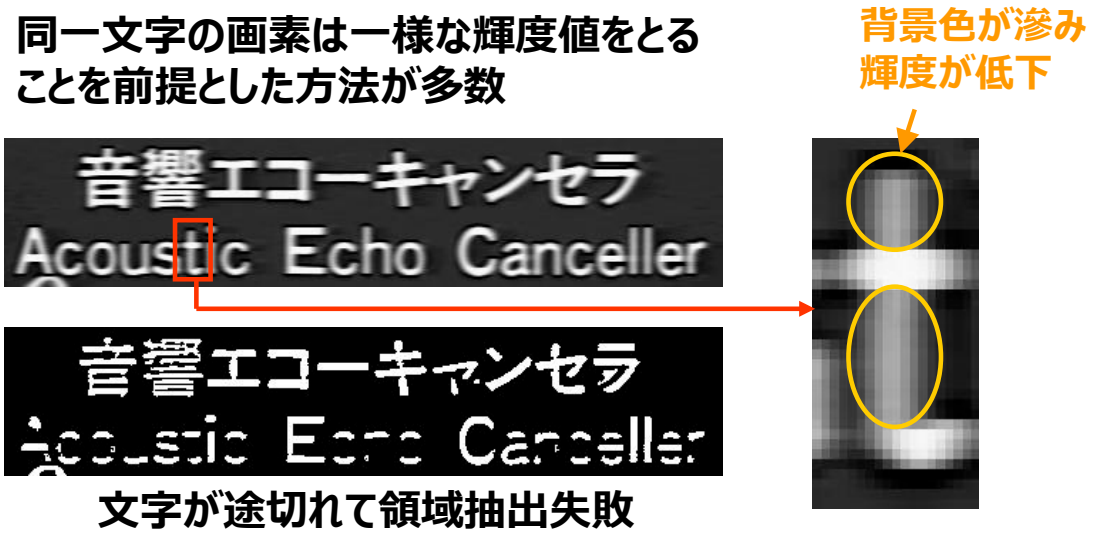


図 4.3.3.2-1　文字領域抽出の従来手法と問題点

4.3.3.3　アプローチ

本研究では、図 4.3.3.2-1 で示したような、文字領域の抽出にあたって、文字が途切れてしまうような従来の問題点を解決する方法を検討する。従来手法で文字部分が正しく抽出できない原因として、放送映像の信号の特性である水平走査ライン上の輝度滲みを詳細に分析する。図 4.3.3.3-1 に示すように、一つ一

つの水平ライン上だけでの輝度分布をみると、輝度劣化した場合でも文字内の輝度変化は小さく、文字と背景の間のコントラストは一定以上の高い場合が多い。また、垂直ライン上だけでの輝度分布をみても同様のことがいえる。

そこで、提案手法では、ライン毎に輝度分布上のコントラストの存在する部分を検出し、ライン毎の検出結果を統合することで、文字領域を形成することとする。これを各水平ライン、各垂直ラインで輝度分布の凸状の部分を取り出し、二値化を行う。2 種類の二値化結果の画像が得られるが、テロップ文字からは面積、形状、位置がほぼ等しい領域が抽出される。一方、テロップ以外の背景からは、必ずしも水平・垂直方向で輝度パターンが凸状になっている訳ではないことから、2 種類の二値化結果は異なるものとなる。

このような文字部と背景部の二値化結果の違いにも着目し、両画像を比較することで。最終的な文字領域を抽出することを試みる。

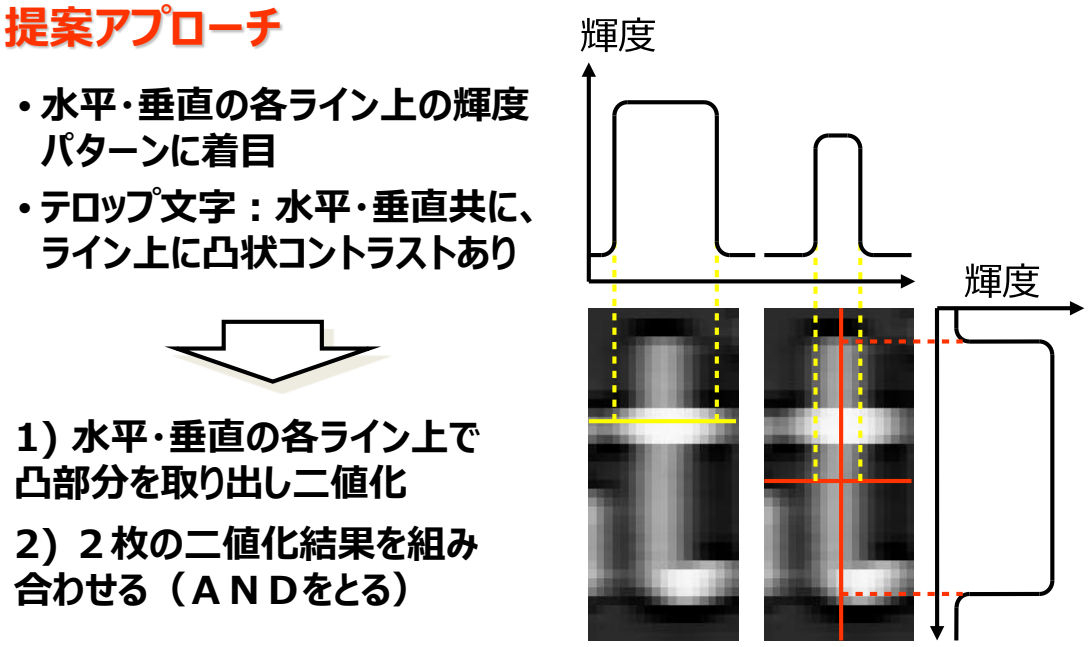


図 4.3.3.3-1　ライン単位の二値化のアプローチ

4.3.3.4　提案方式

提案方式の手順を以下に述べる。

- ✧ 手順 1：　入力テロップ画像の水平ライン毎に輝度分布の二値化を行う（二値画像 A)。
- ✧ 手順 2：　入力テロップ画像の垂直ライン毎に輝度分布の二値化を行う（二値画像 B)。
- ✧ 手順 3：　二値画像 A, B を比較し、両者に面積、形状、位置がほぼ等しく存在する領域を文字領域として抽出し、一方にしか存在しない領域を背景ノイズ領域として除去する。

手順 3 の結果においては、文字サイズが 30 画素×30 画素くらい以下の範囲の比較的低解像度な場合、文字領域の輪郭がギザギザになることがある。これは、手順 1, 2 が水平、及び、垂直の各ライン、すわなち、1 次元の情報をそれぞれ独立に処理するものであるためである。これに対し、2 次元の文字パターンらしさの観点を入れて、ギザギザを平滑化する処理を手順 4 として行う。

- ✧ 手順 4：　手順 3 の結果の文字領域を数画素膨張させた範囲で、輝度ヒストグラムの二値化を行い、最終結果として得る。

手順 4 の二値化の閾値 K は､膨張分の画素だけの輝度値の平均値に比例する値を設定する。これにより、再度、文字ストロークが途切れることを確実に回避しつつ、文字輪郭のギザギザを無くすことを実現する。

4.3.3.5　実験結果と考察

実験結果の例を図 4.3.3.5-1 に示す。(a)は入力テロップ画像であり、画面の中央下部に「画像符号化」「Picture Coding」というテロップが表示されている。(b)と(c)はそれぞれ、(a)の水平ライン単位、また、垂直ライン単位に、輝度パターンの凸状部分を抽出し、二値化した結果である。(b)と(c)を比較すると、文字部分からは、面積、形状、位置がほぼ同じ領域が共に存在するのに対し、背景

部分においては、(b)と(c)で領域の存在パターンは異なる結果が得られている。このため、(b)と(c)の領域 AND を取ることで、最終結果として、文字領域だけが残存する画像(d)が得られる。

また、図 4.3.3.5-2 は輪郭平滑化処理の結果例を示したものである。(a)のライン二値化結果を(b)のように膨張させ、膨張分の画素の輝度平均値をもとに閾値処理をすることで、(c)に示す形で、文字輪郭が滑らかな文字領域が得られた。

以上の処理を約 120 分のニュース番組のテロップ画像 336 枚、総文字数 5263 文字に対して行った結果、全ての文字ストロークを抽出できた文字の割合は、従来の手法では、輝度劣化文字の抽出に失敗し、86.2%であったのに対し、提案手法を用いた文字領域抽出の結果は 94.3%であり、文字領域抽出率が向上した結果として得られた。

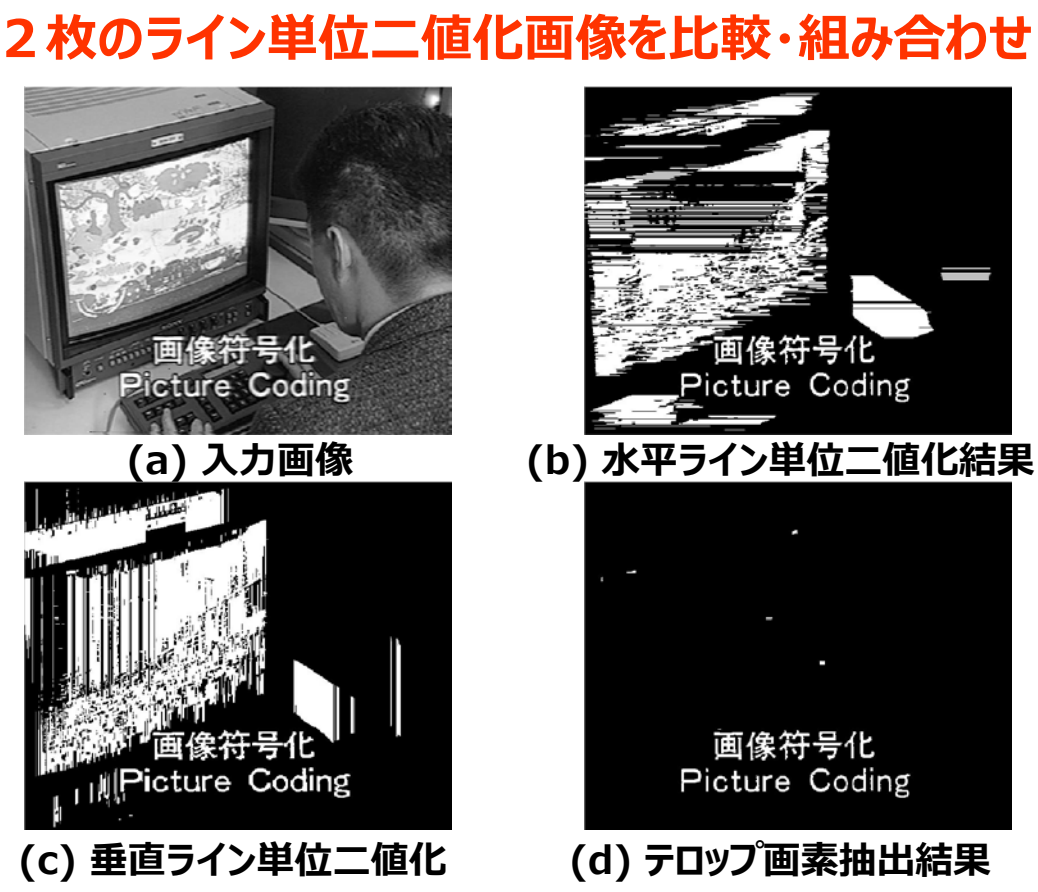
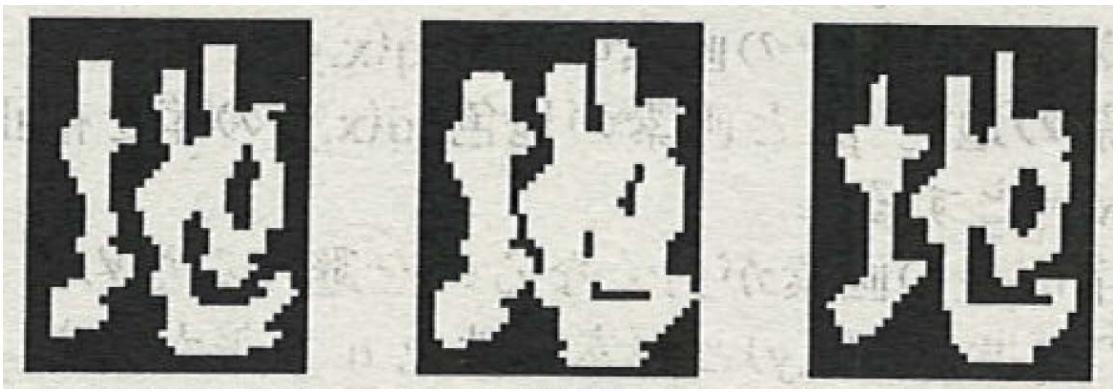


図 4.3.3.5-1　ライン単位二値化の各手順の結果の例



(a) ライン二値化　　(b) 膨張領域　　(c) 輪郭平滑化

図 4.3.3.5-2　文字領域の輪郭平滑化結果の例

4.3.4　テロップ文字列抽出

4.3.4.1　概要

本項では、映像中のテロップ文字列の矩形情報を抽出する方法を提案し、提案手法の実験結果と考察を述べる。映像中に表示されるテロップは映像の意味的な区間、いわゆる話題毎に決まったルールに基づいて表示効果が施されており、映像内容の意味的な構造との相関が高い。

例えば、ニュース映像では、各ニューストピックの最初に見出しを表すテロップが画面の下に他のテロップよりも大きく表示される。大きさ以外にも、見出しのテロップだけ色が異なったり、文字に縁取りが付いていたり、帯状の飾りがついている等、特別な表示効果を施されている場合が多い。

ニュース映像であれば、見出し部分を検出し、次のニュースの見出しまでを一つの区間とすることで、話題単位の抽出が可能となる。このように、テロップには、文字としての言語情報以外に、その表示効果もメタデータを生成する際の手がかりとして利用できる。

筆者らは、映像中のテロップ文字列を囲む矩形の情報（矩形の位置、横幅数、高さ数）を抽出、分類し、映像の意味的な区間を取り出す方法を検討した。本検討では、テロップが表示されている画像から、テロップ文字列の矩形情報を抽出する方法を提案する。

4.3.4.2　従来技術

数種類のニュース映像中のテロップ文字列の矩形、特に、矩形の高さについては、見出しテロップの文字列区系の高さは見出し以外のテロップよりも平均して約 10 画素程度大きいことが分かった。したがって、本検討では、テロップ文字列の高さを数画素程度の誤差内で求める必要がある。

筆者らは、これまでに、テロップ表示画像の二値化結果中の各連結成分の膨張処理に基づいた文字列の抽出方法を提案している[Kurakake]。ただし、映像中には、テロップの他に様々な輝度、色の任意背景が存在するため、[Kurakake]の方法では、二値化結果中に含まれる背景中のノイズ成分も膨張領域に含まれてしまい、実際に文字列範囲よりも大きな範囲で文字列矩形を抽出する傾向があり、要求される精度を満たさない。

また、他にもテロップ文字の輪郭部のコントラストに着目し、エッジ画素の多いラインが集中する部分を文字列範囲とする方法が提案されている[Zhong][Ariki]。ただし、これらの方法も背景から検出されるエッジの影響を受け、文字列の位置、サイズを正確に求まることは困難な場合がある。

すなわち、従来の方法では、文字列矩形の抽出に用いる特徴量に背景部との差を強調することができる「文字列らしさ」が十分に反映されていないという問題があった。

4.3.4.3　アプローチ

本研究では、前項で述べたテロップ領域抽出の二値化処理の結果得られる連結成分とエッジの検出結果を両方を用いることで、背景ノイズの影響を極力抑制した文字列矩形の抽出方法を提案する。提案方法の前提として、予め従来の方法[Kurakake]を使って、一つの文字列のみを囲む仮の文字列矩形画像、及び、文字列の方向情報が抽出されているものとする。この仮矩形内では、各テロップ文字の高さ方向の頂点画素と底辺画素は一直線上に並ぶものと仮定し、提案手法でも[Zhong][Ariki]と同様、文字列方向のライン毎に文字らしさを示す特徴を求め、文字のラインと背景のラインの境界を検出する。

文字列矩形内の文字列のライン上では、文字の連結成分、及び文字輪郭部からエッジ画素が共にほぼ確実に得られる。一方、背景部のライン上では、連結成分、エッジの個数は共に文字部に比べ少なく、また、必ずしも両者は同時に

得られるとは限らない。すなわち、エッジと連結成分は各々単独で扱うと、文字と背景の区別がつきにくい場合もあるが、両者を同時に評価すれば、文字のラインと背景のラインの境界を際立たせることが可能となる。

そこで、本研究では、ライン毎に連結成分の個数とエッジ画素の個数の積を文字列特徴値とし、この値に対し、閾値処理を行うアプローチを取ることとする。

4.3.4.4 提案方式

本研究で提案する画像中の文字列矩形の抽出の手順を以下に述べる。

- ✧ 手順 1： 仮の文字列矩形内で文字列方向のライン毎に文字列特徴値を求める。
- ✧ 手順 2： 文字列特徴値の上位 M 個の平均値 Ave を求める。
- ✧ 手順 3： 閾値 Th = Ave / N 以上の文字列特徴値を持つラインを文字列ラインとして残す。
- ✧ 手順 4： 求めた文字列ラインの最高、最低のラインの間の距離を文字列の高さ、あるいは、文字列の幅の値として求める。

なお、以下の実験では、文字列特徴値の計算に使用するエッジ画素の検出には、文字らしい部分から選択的に得られるエッジペア画素を使用する。また、閾値処理では、M = 10, N = 10 とした。

4.3.4.5　実験結果と考察

提案手順の実データへの適用結果を述べる。図 4.3.4.5-1 に結果の例を示す。

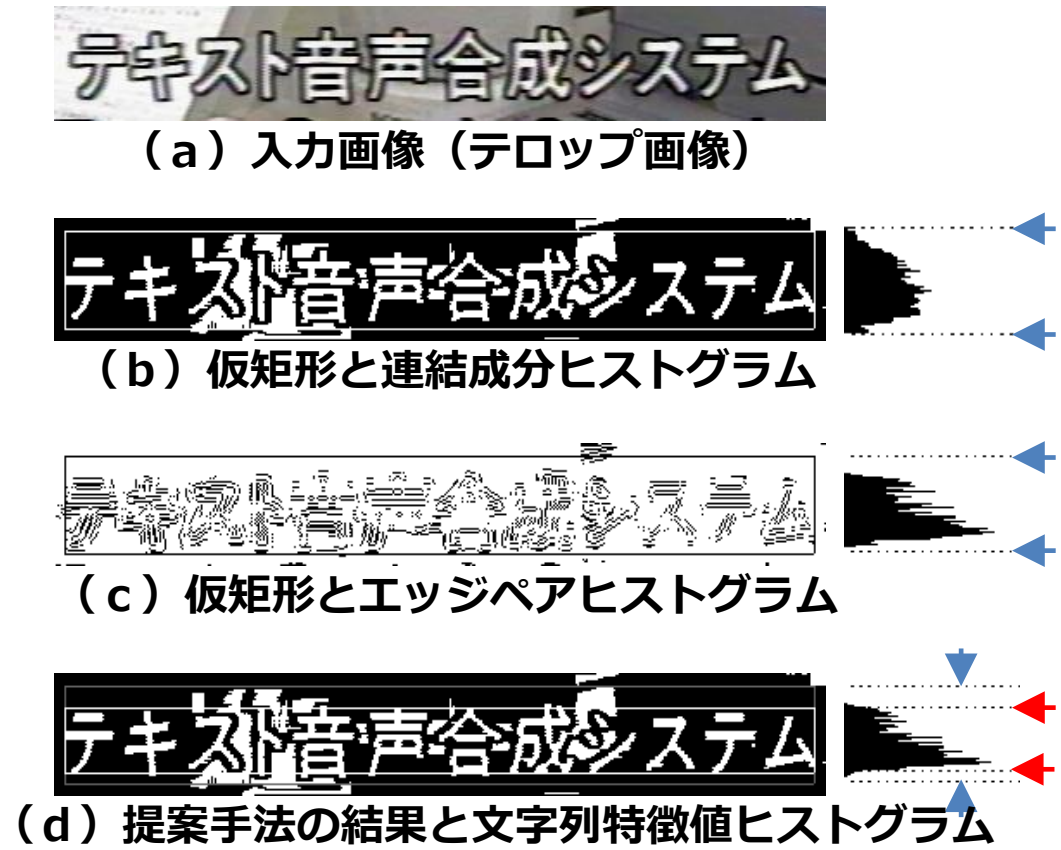


図 4.3.4.5-1　テロップ文字列矩形抽出の実験結果の例

図 4.3.4.5-1 の(a)は入力画像である。(b)は[Kurakake]の方法を用いて得た文字列矩形の範囲、及び、文字列方向のライン毎に連結成分の個数をカウントしたヒストグラムである。(c)は 4.3.2 項で述べたエッジペア検出結果、及び、エッジペアのヒストグラムである。(b)では背景ノイズの影響、(c)では文字の縁取り部分からのエッジの影響で文字ラインと背景ラインの区別がつきにくい。(d)は文字列方向のライン毎に得た文字列特徴値のヒストグラム、及び、閾値処理により得られた文字列矩形を示す。提案した文字列特徴値により、文字のラインと背景のラインの差が強調されたため、(b)の矩形よりも正確に文字列の上限と下限を抽出できた。

ニュース映像中のテロップ表示画像 140 枚に対し、提案手法を用いて全画像中のテロップ文字列の高さ（縦書きの文字列の場合は幅）の値を抽出したところ、図 4.3.4.5-2 に示すように、双峰性の分布が得られた。この分布を二分割し、高さ値の大きい方の山に含まれる画像を見出しテロップ画像として検出したところ、各ニュートピックの冒頭に表示される見出し画像 21 枚を全て正しく検出でき、提案手法の有効性が確認できた。

今後は、他のテロップ属性の抽出法、及び、ニュース以外の映像への適用を検討していく。

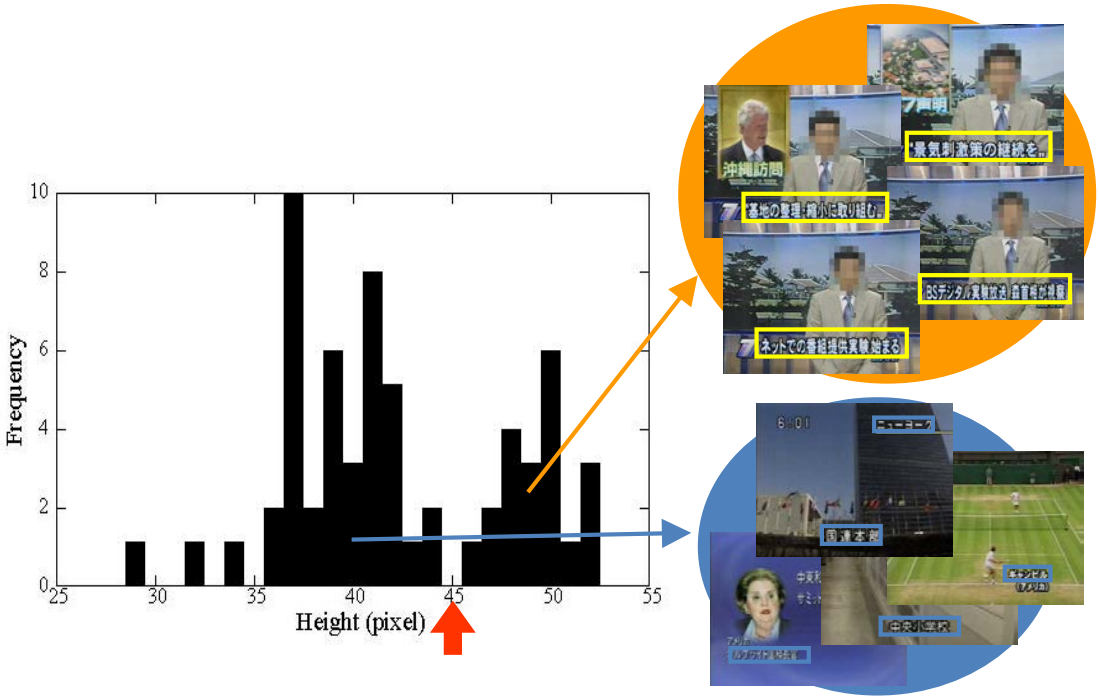


図 4.3.4.5-2　ニュース番組全体のテロップ画像からの見出しテロップ抽出結果

4.4　テロップ文字の視覚特徴に基づく区間メタデータ自動生成

4.4.1　概要

本節では、前節 4.3 で述べた、テロップ文字の認識処理の中でも、特に、テロップ文字列抽出の処理で得られるテロップ文字の大きさ、画像内の表示位置といった視覚的な特徴に基づく区間メタデータの生成方式を提案し、実験結果と考察を述べる。テロップ文字の視覚特徴の他、カット点、カメラワーク、人物の動作検出を組み合わせ、従来技術では困難であった、ニュース番組中のニュースストピックの冒頭の開始タイミング等の意味的なシーンだけを選択的に検出し、区間メタデータの生成作業のコストを低減させる情報を得る方式として提案する。

以降、本節では、4.4.2 項で提案手法のアプローチ、4.4.3 項でニュース番組を対象とするメタデータ生成方式、4.4.4 項で野球中継番組、また、4.4.5 項でサッカー中継番組を対象とする区間メタデータの生成方式を述べる。4.4.6 項で、実験結果と考察を述べる。

4.4.2　アプローチ

テロップ文字の視覚特徴としての文字の大きさ、画像中の表示位置の情報に基づき、区間メタデータを生成する方式のアプローチを述べる。テロップ文字の視覚特徴だけでも、ニュースストピックの開始時間や野球中継の得点シーンのタイミングを捉えることはできるが、シーン視聴サービスの区間メタデータとして、シーンの開始時間、終了時間を作り出すために、テロップ以外の映像特徴も組み合わせた方式を検討する。

組み合わせ方として、メディア解析処理により得られる各種映像イベントの発生タイミングの順番や発生間隔の時間に対して条件（ルール）を設け、条件を満たす場合、区間メタデータとなる開始時間と終了時間として、どのタイミングを定義するかを規定するアプローチを取る。

例えば、ニュース番組の場合、ニューススタジオのシーンに切り替わる際のカット点やその後すぐに表示されるニュースストピックのテロップが表示される

タイミング等に対する表示順や発生間隔に条件を設定する。このようなアプローチを取ることで、従来は番組中の全てのカット点から区間メタデータに相応しいカット点を探さなければいけなかったのに対し、区間メタデータの候補だけを検出することができる。これは、メタデータ生成のコストを圧倒的に下げるものと考えられる。

4.4.3　ニュース番組向けの区間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ーストピックの区間メタデータの生成ルールを策定する。区間メタデータの生成ルールは、図 4.4.3-1 に示すように、各ニュースト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、その近傍のカット点の検出結果と組み合わせて定義する。具体的な区間メタデータの生成ルールは以下とする。

- ✧　開始時間：ニュースタイトルのテロップの直前のカット点の時間
- ✧　終了時間：次のニュースタイトルのテロップの直前のカット点の時間

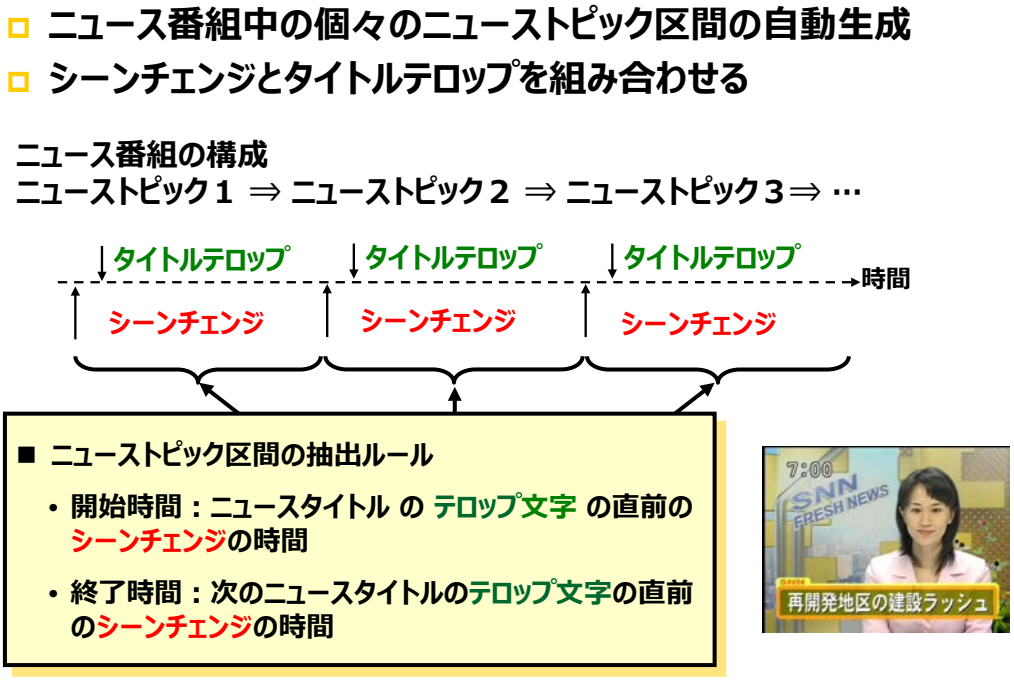


図 4.4.3-1　ニューストピックの区間メタデータの自動生成ルール

4.4.4 野球中継番組向けのメタデータ生成ルール

野球番組のダイジェスト映像には、得点シーンやファインプレーシーンなどが含まれる場合が多い。例えば、得点シーンにおける映像シーンの流れ、演出としては、一般に、ピッチャーが投球し、その投球に対し、ホームラン等の事象が起こり、ホームインして得点のテロップが表示される。得点シーンを定義する区間メタデータとしては、この一連の流れにおける映像シーンの開始時間と終了時間が相応しいと考えられる。

本研究では、野球中継番組用のメタデータとして、得点シーンに関する区間メタデータの自動生成ルールを策定した。図 4.4.4-1 に示すが、開始時間、終了時間は、それぞれ以下の通りである。

- ✧ 開始時間：得点テロップの直前の投球動作シーンの開始時間
- ✧ 終了時間：得点テロップの表示終了時間

開始時間としてのピッチャーの投球動作については、[Fujii]の方法を用いて自動抽出可能である。[Fujii]の方法は図 4.4.4-1 内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い、マッチ度が極大になる時間を検出する。

また、得点数字のテロップ文字については、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、前節 4.3.4 で説明した方法を用いて、選手名等の他のテロップと区別して検出できる。以上、投球動作検出、得点テロップ検出という 2 種類のメディア解析の結果を組み合わせることで得点シーンの区間メタデータを検出する。

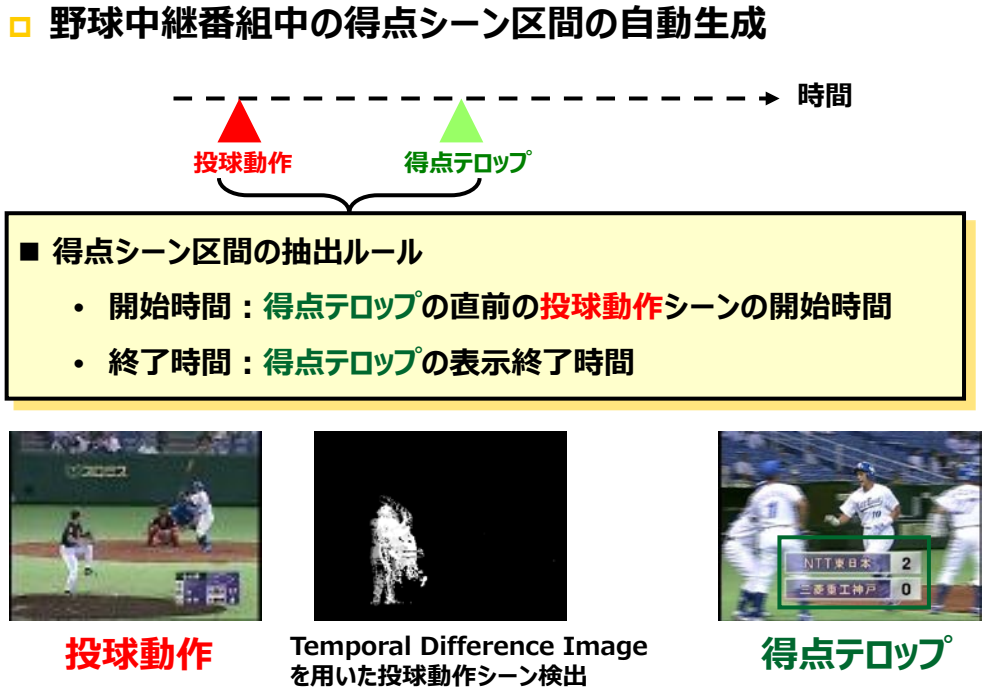


図 4.4.4-1　野球中継番組の得点シーンの区間メタデータ生成ルール

4.4.5　サッカー中継番組向けのメタデータ生成ルール

サッカー中継番組におけるシーン単位のダイジェスト視聴においては、試合中の盛り上がるシーンとして、ゴールシーンやシュートシーンが含まれる場合が多い。例えば、シュートシーンにおける映像シーンの流れ、演出としては、一般に、シュートが打たれる前に、視聴者がボールの動きがよく確認できるようゴール付近を中心として、ズームやパン、チルトといったカメラワークが存在し、シュートが打たれると、シュート後にシュートを打った選手の名前がテロップで表示される。また、シュート後には、リプレイ映像が流されることが多いが、リプレイ映像の開始時に映像演出として、特殊な CG パターンが用いられる場合もある。シュートシーンを定義する区間メタデータとしては、このカメラワーク、テロップ、CG パターンの一連の流れにおける映像シーンの開始時間と終了時間が相応しいと考えられる。

本研究では、サッカー中継番組用のメタデータとして、シュートシーンに関する区間メタデータの自動生成ルールを策定した。図 4.4.5-1 に示すが、開始時間、終了時間は、それぞれ以下の通りである。

✧ 開始時間：カメラワーク（ズームイン）の開始時間

✧ 終了時間：CG パターンが表示される直前の時間（リプレイ映像に入る直前のタイミング）

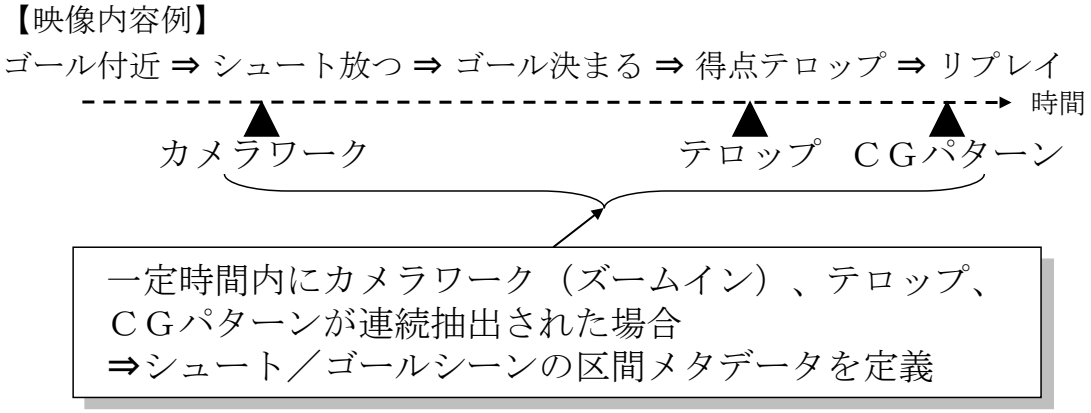


図 4.4.5-1　サッカー中継番組のシュートシーン／ゴールシーンの区間メタデータ生成ルール

4.4.6　実験結果と考察

メディア解析結果へのルール適用によるメタデータ自動生成方式の精度評価を行った。ニュース番組 3 本、野球中継番組、サッカー中継番組各 1 本に対して、それぞれ 4.4.3 項、4.4.4 項、4.4.5 項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評価した。表 4.4.6-1 に結果を示す

全ての番組コンテンツに対し、再現率、適合率とも一定以上の値が得られ、複数のメディア解析結果の組み合わせに対するルール適用によるメタデータ生成の効果が確認できた。特に、ニュースに関しては、再現率、適合率とも 100%であり、自動生成された区間メタデータを手作業で修正する必要なく、そのままサービス利用できる品質のものが得られたといえる。各ニューストピックの冒頭のタイトル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、サッカーについては、ほとんど漏れなく所望のシーンが得られたが、ニュースに比べ誤抽出が多かった。

表 4.4.6-1　提案方式による区間メタデータの生成精度

	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 各ニューストピック	100%（= 34 / 34）	100%（= 34 / 34）

今回検証した 4.4.3 項、4.4.4 項、4.4.5 項で述べた各ルールは、全て番組放送時の映像に含まれるテロップの情報を有効に活用したものである。ただし、各ルールはテロップを活用するという意味では共通であるものの、ニュース番組、野球中継番組、サッカー中継番組といった番組ジャンルの違いによらない汎用的なルール作りを指向したものではない。

4.4.3 項、4.4.4 項、4.4.5 項で述べた各ルールは、例えば、映像制作時にテロップを挿入する以前の素材映像に対しては効果を発揮しにくいルールであり、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。勿論、汎用性の高いメタデータ生成方式を確立することが重要ではある。しかしながら、実際の映像視聴サービス向けのメタデータ生成作業としての運用時においては、映像にテロップが入っているかどうか以外にも、例えば、ニュース番組のアナウンサーの読み原稿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、最終的にどのようなサービスに利用するのかによって、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ことから、汎用性の高い方式を検討するのは勿論、それだけでなく、メディア認識以外にも GUI、システムハードウェア構成も含め、メタデータ生成方式のカスタマイズの容易性を高めていくことが重要と考える。

4.5 Telop on demand システム

我々は、テレビ放送映像全チャンネル分を録画蓄積しながら、並行して、リアルタイムに、4.3 節で述べた、テロップ文字の認識処理を実行するシステム「Telop on demand」を開発した。Telop on demand は、ユーザ向けに、録画した映像のシーン全体閲覧と個別シーン検索の機能を提供する。

テロップ文字が表示されるフレーム画像「テロップ画像」と、テロップ画像中の文字認識結果と、文字の大きさと表示位置からなる属性抽出結果は、Telop on demand のメタデータデータベースに格納される。

Telop on demand は、テロップ文字の認識に関する一連の処理が実装されている。図 4.5-1 と図 4.5-2 には、Telop on demand のユーザインタフェース機能の画面例を示す。テロップ文字認識の結果として得られるテロップ文字のテキスト情報やテロップ文字の大きさ、表示位置の情報を活用して、以下のユーザインタフェース機能を提供する。

テロップ文字という映像の意味内容を直接反映した情報を手がかりに、映像全体のブラウジングやキーワードでのシーン検索を実現した。

- ブラウジング： ニュース番組の各ニューストピック全体を一画面で俯瞰閲覧できる。具体的には、図 4.5-1 の左側に示すように、各ニューストピック冒頭のタイトルテロップ文字の画像を一覧できる。また、タイトルテロップ文字の画像を押下すると、図 4.5-1 の右側に該当する各ニュース映像区間に含まれる、その他のテロップ文字も含め、テロップ文字画像が一覧できる。ユーザは簡単な操作で、ニュース内容の詳細を把握することができる。
- 検索： テロップ文字の認識結果のテキストをデータベース化しておくことで、ユーザはシーン映像のキーワード検索ができる。具体的には、放送番組全チャンネル分の映像から、コマーシャル映像を含め、番組横断的に、所望のシーン映像の検索ができる。キーワードがテロップ文字として含まれているテロップ文字画像が検索結果として表示されるが、図 4.5-2 は、映像中に「NTT」というテロップ文字が含まれるシーンを検索した例である。このように、ユーザはテロップ文字画像を押下することで、該当画像を先頭とする映像を再生視聴することができる。

図 4.5-1　Telop on demand のユーザ向け画面例①

図 4.5-2　Telop on demand のユーザ向け画面例②

4.6 まとめ

本章では、メタデータ生成の作業コストの低減という本研究の課題に対し、特に、テロップ文字を活用し、区間メタデータを自動的に生成する方法を中心に述べた。

最初に、テロップ文字認識、あるいは、テロップ文字認識以外のメディア解析について、従来技術を述べた。従来技術の取り組みにおいては、メタデータ生成の作業コストの削減というレベルでのテーマ設定は存在せず、いずれも、メディア解析技術そのものの精度評価を行うものであった。また、従来技術は、映像全体の中の全てのテロップ文字を対象とする技術であり、ニュートピックの開始タイミングという映像の意味内容を反映した区間メタデータの生成方式は国内外の著名な学会においても発表は存在しなかったことを述べた。このようなことから、本研究の課題であるメタデータ生成の作業コストの削減に対し、特に、区間メタデータの生成について、従来技術を適用しても解決すべき、大きな問題があることを述べた。

次に、テロップ文字と映像内容の意味内容の相関関係に着目し、区間メタデータ、意味メタデータの自動抽出を想定したテロップ文字の認識方式を述べた。テロップ画像検出、テロップ領域抽出、テロップ文字列抽出を順に述べ、それぞれ、従来技術の問題点を指摘し、それを解決する方式として本研究の提案方式を述べた。テロップ画像検出では、エッジペア特徴を提案し、従来技術よりも、誤検出を抑制した方式を提案、評価結果を述べた。テロップ領域抽出では、水平ライン、垂直ライン単位で二値化を行い、組み合わせることで、従来技術では抽出が困難であった、同一文字内で輝度劣化がある場合でも正しく領域抽出できる方式として提案、評価結果を述べた。テロップ文字列の抽出では、エッジペアに基づき、従来技術よりも、高精度に文字列の大きさを抽出する方式を提案し、ニュース番組中のニュートピックの開始タイミングだけを正しく抽出することに応用できることを示した。

また、テロップの大きさ、表示位置という視覚特徴に基づき、カット点や人物動作等の他の映像特徴と組み合わせた区間メタデータの自動生成方式を提案した。各映像特徴の発生タイミングの順番や発生間隔の時間に対して条件（ルール）を設けるアプローチを示した。このアプローチにより、ニュートピック、野球の得点シーン、サッカーのシュートシーンの開始時間と終了時間を自動生成する方法を提示し、実験結果と考察を述べた。ニュートピックの開始時間については、抽出率 100%という形で、ほぼ人手作業による修正の手間が不要となるレベルで抽出できることを示した。

最後に、応用システムの例として、テレビ番組を全チャンネル 24 時間分の映像データに対し、テロップ文字をよる映像ブラウジング、キーワード検索のユーザインタフェースを提供する Telop on Demand システムを紹介した。カット点等の信号レベルに近い情報に比べ、テロップ文字は映像の意味内容を示すことから、従来の映像配信システムに比べ、映像内容の短時間に理解する効果があるシステムとして実装したことを示した。

なお、本章で述べたテロップ文字の認識方式については、テレビ放送信号に対しての処理であるが、研究当時の 2007 年は、現在のデジタル形式ではなく、アナログ形式であった。テロップ文字の認識方式のうち、テロップ領域の抽出については、画像中の水平ラインに沿った輝度値の滲みにより生じる同一文字パターン内の輝度劣化に対応する方式として提案した。これは、アナログ形式の放送信号特有の問題であり、研究当時としては、従来技術の問題を解決する非常に有用性がある方式であった。しかしながら、2018 年現在の輝度劣化のないデジタル形式の放送信号においては、オーバースペックな方式になると考えられる。ただし、提案方式の水平、垂直の各ライン上で凸状の輝度パターンを示すという特徴は変わらないことから、現在においても適用可能な方式と考えられる。放送事業者には、現在も、過去の放送番組映像がアナログテープ形式のライブラリとして残っていることから、本研究のテロップ領域抽出の方式は、このような過去のアナログ形式の放送番組、また、最新のデジタル形式の放送番組の両方に対応できる方式として適用可能なものと考えられる。

第5章 制作済番組向けメタデータ生成の作業コスト削減

5.1 概要

本章では、第 4 章で述べたテロップ文字認識によるメタデータ候補の自動生成を含め、メタデータ生成の作業者によるメタデータ生成の作業コスト削減に関する作業モデルを提案する。また、提案作業モデルによるメタデータ生成と人手作業によるメタデータ生成の作業時間に関する比較実験の結果、及び、考察を述べる。本章では、一度、放送した番組映像をネット配信向けに 2 次利用するケース、すなわち、制作済番組を対象とした検討を行う。

メタデータ生成の作業コスト削減効果の評価のために開発したメタデータ生成・編集システム「SceneCabinet / NBS」は、第 4 章で述べたメディア解析技術により生成されたメタデータを一覧表示する等、3.1.4 項で述べたメタデータ生成向けのユーザインタフェースシステムの要件を満たす機能を備えている。図 5.1-1 に該当の要件を再掲する。

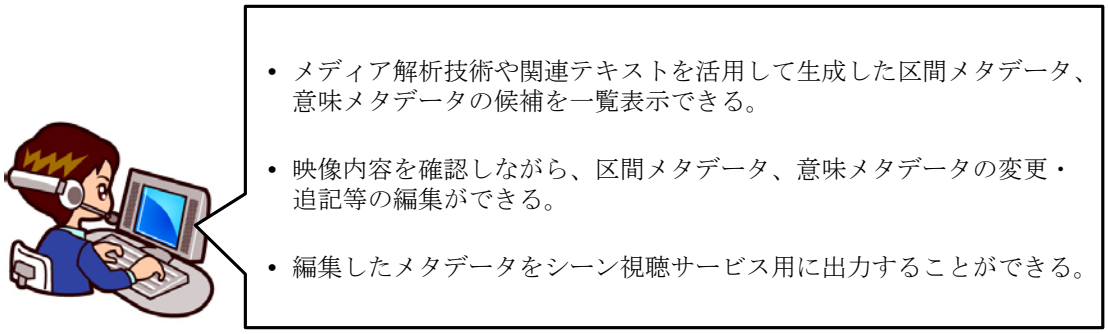


図 5.1-1　ユーザインタフェースシステムの要件例

SceneCabinet / NBS の機能の具体例として、映像中のカット点、カメラワーク、音楽等、シーン情報を一覧表示するキー画像ブラウザ、及び、映像中のテロップ文字や音声の認識結果に対して自然言語処理を適用することで、シーンのタイトル、概要文、キーワードといった情報を自動生成、編集するメタデータエディタ等の機能を備えている。

更には、番組の台本テキストを読み込んで、キーワード等のメタデータを自動抽出する機能も備えている。すなわち、メタデータを位置から手入力するのではなく、メディア解析処理の結果や既存のテキスト情報を再利用することで、手作業にかかる作業時間の短縮化を図るためのユーザインタフェースである。

以降、本章では、5.2 節では、SceneCabinet / NBS の機能概要、及び、メタデータ生成の作業コスト削減を実現するための作業モデルについて述べる。5.3 節では、制作済のニュース番組、サッカー中継番組に対するメタデータ生成作業について、SceneCabiet / NBS を用いた作業モデルを述べる。5.4 節では、提案する作業モデルによるメタデータ生成の低コスト化効果の検証実験の結果と考察を述べ、5.5 節で本章のまとめを述べる。

本章の内容は、筆者らの公表論文[J1],[C4]に基づくものである。

5.2　メタデータ生成システム「SceneCabinet / NBS」

5.2.1　機能概要

メタデータ生成用のユーザインタフェースシステム「SceneCabinet / NBS」の機能概要を述べる。なお、SceneCabinet は筆者らの研究プロダクトのシリーズ名称であり、NBS は ”Next-generation Broadcasting System”の略である。前世代システム SceneCabinet [Taniguchi] との機能差分を含め、SceneCabinet / NBS が備える主な機能を以下に示す。

SceneCabinet / NBS の主な機能 (a)～(e) :

(a) 映像・音声データ、関連テキストデータが入力できる。

(b) 第 3 章で述べた、メディア解析技術を活用した区間メタデータ、意味メタデータの自動生成ができる。

(c) メタデータ自動生成の結果の全体一覧表示、個々のメタデータの内容確認、編集ができる。

(d) 区間メタデータ、意味メタデータを一つのファイルとして出力できる。この際、フォーマットは国際標準の TV-Anytime 形式に準拠したものを出力できる。

(e) 以上の機能は Windows パソコン上に実装されている。

図 5.2.1-1 に SceneCabinet / NBS の全体構成を示す。映像、音声データ、及び、関連テキストデータを入力とし、TV-Anytime 形式の区間メタデータと意味メタデータを一つのファイルとして出力するものである。SceneCabinet / NBS 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」から構成される。メタデータ生成の作業者は、入力データを読み込ませ、各機能を利用し、メタデータを出力する一連の作業を行うイメージである。以下、メタデータ生成の作業コスト削減を目的とした SceneCabinet / NBS の各機能部を詳細に述べる。

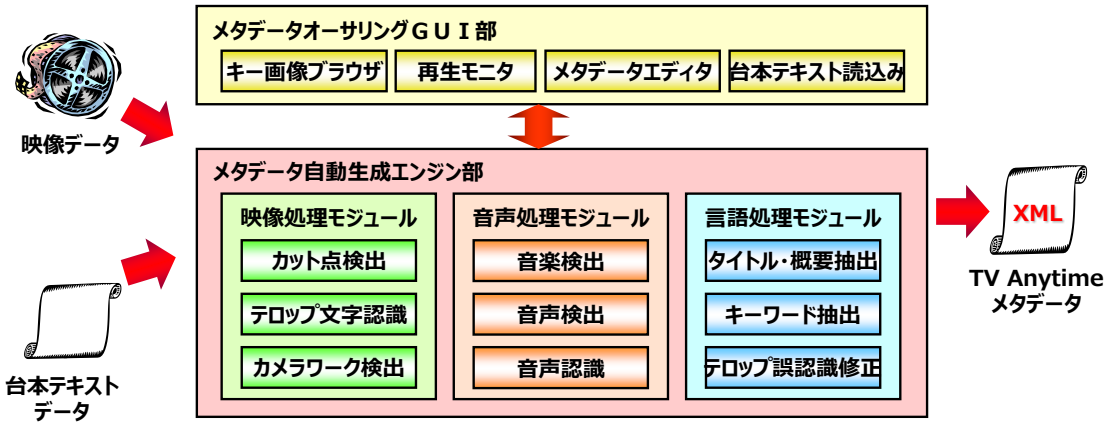


図 5.2.1-1： SceneCabinet の機能構成

5.2.2　メタデータ自動生成エンジン部

5.2.2.1　メディア解析機能

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、及び、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、各認識結果をテキストデータとして出力する。また、自然言語処理により、テロップ文字、音声の認識結果から映像シーンのタイトル、概要、キーワードを自動抽出する。更には、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る。

5.2.2.2　メタデータ生成ルールのカスタマイズ機能

SceneCabinet / NBS は、各メディア解析エンジンを組み合わせてメタデータを生成する際のルール記述をカスタマイズする機能も備える。図 5.2.1-1 に示すように 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 PC 上で同時実行すると、必要以上の情報が取得され、システム負荷も高くなる。例えば、4 章で述べたように野球中継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、他のメディア解析エンジンの結果は必須ではない。SceneCabinet / NBS では、番組ジャンルに合わせて、必要なメディア解析エンジンの ON/OFF のパターンを定義できるよう実装した。更には、区間メタデータの自動生成を目的とした、メディア解析結果の発生順序に関する条件も設定ファイルに記載できるように実装した。これにより、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、また、区間メタデータ生成の新しいルールを自由に追加・カスタマイズできる。

5.2.3　メタデータオーサリング GUI 部

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、最終的な区間メタデータ、意味メタデータを完成させるための作業者向け GUI 機能を提供する。メタデータ生成の作業コスト削減には、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る。図 5.2.3-1 にメタデータオーサリング GUI 部の画面例を示す。以降、各 GUI パーツの機能を詳細に述べる。

図 6.2.3-1：メタデータオーサリング GUI 部の画面例

5.2.3.1　キー画像ブラウザ

キー画像ブラウザは「メタデータ自動生成エンジン部」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閲覧できる。各画像はカット点、テロップ文字など、メディア解析の処理の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る。

5.2.3.2　メタデータエディタ

メタデータエディタでは、編集対象の映像シーン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意味メタデータ（タイトル、概要文、キーワード)、及び、代表画像を自動抽出し、その結果を確認、編集できる。

5.2.3.3　再生モニタ

再生モニタは、キー画像ブラウザ中のサムネイル画像を選択すると、その画像に対応する時間からすぐに映像を再生させることができ、かつ、映像中のフレーム画像を 1/30 秒や 10 秒といった、様々な粒度の時間間隔で移動させ時間情報を効率的に微調整することができる。前述のキー画像ブラウザと再生モニタを利用することで、区間メタデータを簡単に確認、修正することが可能である。

さらに、再生モニタには、音声レベルの波形表示機能も付いており、一定の時間区間の音声の有無を映像再生することなく確認でき、音声の有無を考慮した区間メタデータの確認・設定作業が効率的に実施できるようになっている。

5.3　SceneCabinet / NBS を用いたメタデータ生成の作業モデル

5.3.1　区間メタデータの生成作業

区間メタデータの生成作業は、メタデータを付与したいシーンの開始時間と終了時間になるべく近い時間情報が自動的に取得できると、作業時間が短く済む。

そこで、人手作業の前段で行うメディア解析によるメタデータ候補の自動生成処理においては、SceneCabinet / NBS のメディア解析エンジン部に実装した[Taniguchi],[Minami][C2]で提案されているメディア解析技術により、映像中の

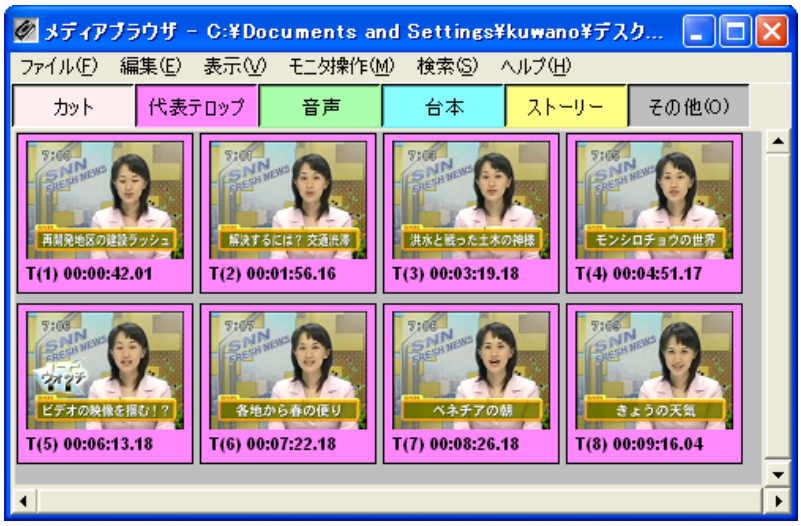
カット点、テロップ文字、カメラワーク、音声、音楽といった映像中のポイントとなるシーン情報を区間メタデータの手がかりとして生成する。

本処理部においては、特に、従来システムには存在しない新たな機能として、ニューストピック情報など、映像中の意味レベルの情報を取り出す映像認識の機能を実装した。具体的には、映像中のテロップ文字について、単に映像中の全てのテロップ文字を区別なく抽出するのではなく、テロップ文字の大きさ、画像中の表示位置について、予め指定した条件を満たすものを代表テロップとして他のテロップと区別して抽出する機能を新たに追加した。

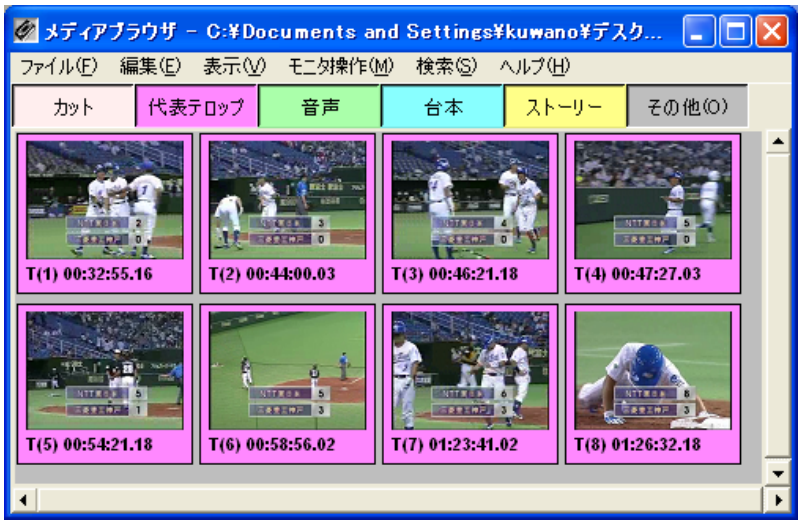
テロップ文字の大きさ、画像中の表示位置を抽出する技術としては、前章で述べた画像の 2 値化結果と輝度エッジの検出結果を組み合わせて利用する方法[C2]を用いた。これにより、例えば、ニュース番組中の各ニューストピックの冒頭に表示されるニュースタイトル表示や、サッカー中継におけるシュートシーン後の選手名やゴール後の得点表示といった番組内で毎回同じ位置に表示され、意味的なシーンと関連性の高いテロップを他のテロップと区別して抽出することができる。

図 5.3.1-1 に代表テロップの例を示す。これらのテロップの抽出結果は、ニューストピック、得点シーン、シュートシーンといった意味レベルの情報、すなわち、最終的に生成したい区間メタデータに近い情報として有効に利用することができる。

メタデータ生成の作業者は、SceneCabinet / NBS のオーサリング GUI のキー画像ブラウザ上で、代表テロップのサムネイル画像を一覧表示し、その時間情報を再生モニタにおいて微調整することで、区間メタデータを完成させる。5.2.3.3 で述べた、再生モニタ上の様々な時間粒度の映像ジャンプボタン、音声波形表示などを駆使することで、音声の有無を映像再生することなく確認でき、音声の有無を考慮した区間メタデータの設定作業が効率的に実施できるようになっている。



（ａ）ニュース番組中のニュースタイトルのテロップ
(a) Titles of each topic in a news program



（ｂ）野球中継の得点数字のテロップ
(b) Scores in a baseball game program

図 5.3.1-1　代表テロップの例

5.3.2　意味メタデータの生成作業

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る。そこで、事前に作成された区間メタデータに対応する映像区間に対し、テロップ文字認識[C2]、音声認識[Ohtsuki]の技術を利用して、映像区間中に含まれるテロップ文字、音声の認識結果などのテキストの中から映像シーンの意味メタデータの自動抽出を実施する。

意味メタデータの自動抽出には[Hayashi]の自然言語処理を利用する。なお、[C2][Ohtsuki]の技術はそれぞれニュース番組に適用した場合、テロップ認識精度は約 80%、アナウンサーの発話内容についての音声認識精度は 90%以上の性能である。図 5.3.2-1 に意味メタデータの自動生成の概念図を示す。入力テキストに対して、処理ルールを適用させることで、映像シーンのタイトル、概要文、キーワードを抽出するものである。

以下に、ニュース番組の各ニューストピックに対する意味メタデータ生成ルールの例を示す。

(1) タイトル:　代表テロップ(ニュースタイトルのテロップ文字(図 5.3.1-1))の認識結果を設定

(2) 概要文：　代表テロップが検出された時間を含む一定時間内の音声（ニュース冒頭にアナウンサーが話すニュースサマリと仮定）の認識結果を設定。

(3) キーワード：　ニューストピック区間内の全てのテロップ認識結果、音声認識結果に含まれる重要単語を設定。

本機能は、SceneCabinet / NBS のメタデータエディタ上の意味メタデータ抽出ボタンを押下することで実行できるよう実装した。このため、作業者は、一旦、区間メタデータを生成後、意味メタデータ抽出ボタンを利用し、メタデータエディタ上に提示される自動生成結果を目視で確認し、必要に応じてその内容を編集するだけの作業を実施するだけで良い。

従来システムでは、意味メタデータの自動抽出機能そのものが備わっていないため、作業者は意味メタデータのテキストを一から手入力する必要があったが、提案作業モデルによれば、作業者は簡単な確認作業を実施するだけでよく、作業時間も短く済ませることが可能となる。

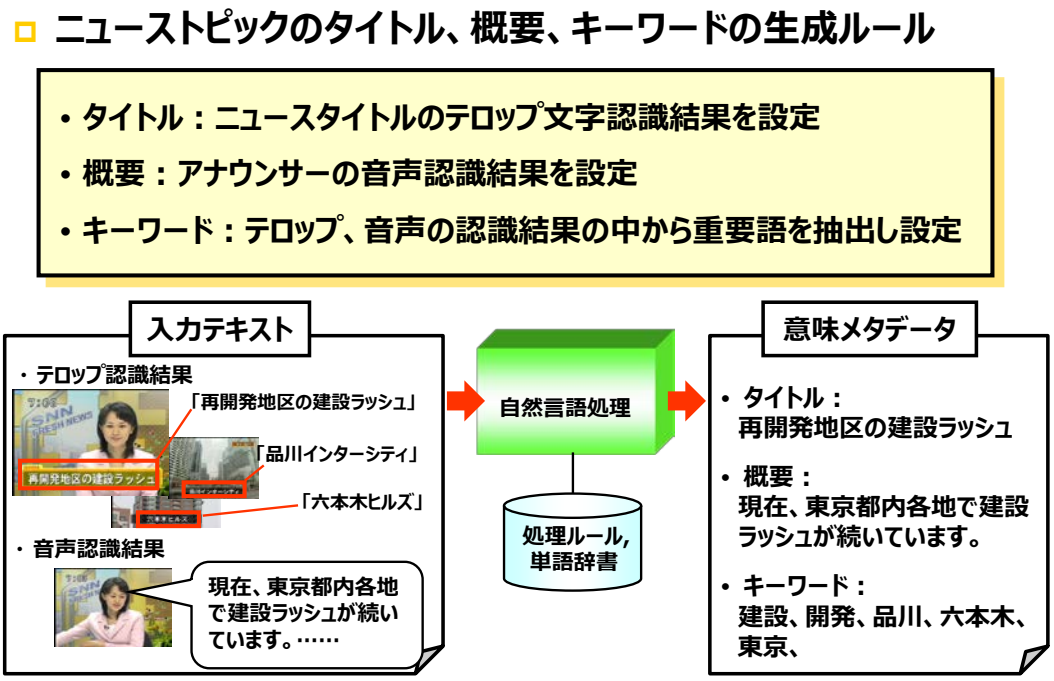


図 5.3.2-1　ニューストピックの意味メタデータの自動生成ルール

5.3.3　台本テキストの利用

番組制作時に作成される台本テキスト情報もメタデータ生成に再利用することが可能である。本研究では、台本テキストとは、①番組中のシーン毎の開始、終了の予定時間、②各シーンのセリフ内容や状況を説明するテキストが含まれるものをさすこととする。

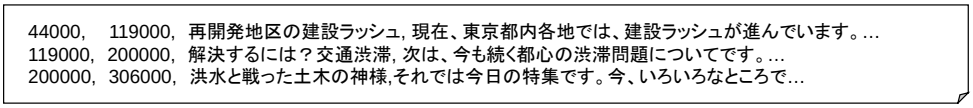
例えば、ニュース番組制作時に作成される各ニューストピックの放送予定時刻とアナウンサーの読み原稿、サッカーや野球などのスポーツ番組におけるシュートシーン、あるいは各バッターシーンなどの個々のプレー内容と試合中の時間情報が記載されるスコアシートなどがそれにあたる。

前述の①は区間メタデータ、②は意味メタデータを生成する上で有効活用することが可能である。具体的には、区間メタデータの生成は、カット点やテロップ、音声区間といった映像・音声認識結果の情報だけでなく、台本中の時間情報も手がかりの一つとして利用する作業内容になる。

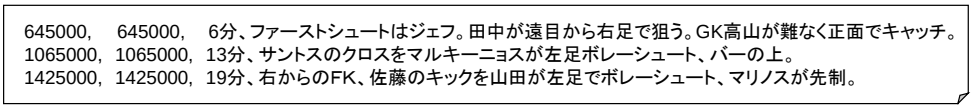
SceneCabinet / NBS は、シーン毎に開始時間、終了時間、説明テキストの順に記載される CSV 形式の外部ファイルとして、台本テキストを読み込み（図

5.3.3-1)、その中の各シーンの時間情報に対応する映像中のサムネイル画像をキー画像ブラウザに表示することが可能である。

本機能と再生モニタを利用して、区間メタデータが決定できる。また、意味メタデータについては、テロップや音声の認識結果に加え、台本中のテキスト情報を利用することで、より精度の高い情報を自動抽出できるようになることが見込める。



（a） ニュースの読み原稿
(a) Script of news topic



（b） サッカーのスコアシート
(b) Score sheet of soccer game

図 5.3.3-1： 台本テキスト情報の例

5.4 メタデータ生成の作業コスト評価実験

5.4.1 実験概要

区間メタデータと意味メタデータを生成する作業について、SceneCabinet / NBS を用いる提案方式と従来方式で行った場合の各作業モデルの作業時間などの比較実験を行い、提案方式のメタデータ生成の効率化効果に関する検証を行った。

実験では 4 人の被験者を対象に以下の 3 種類の作業モデルの作業時間の比較、及び、作業の疲労感などに関するアンケートを実施した。

(1) 提案モデル A：SceneCabinet / NBS を利用（関連テキスト無し）

(2) 提案モデル B：SceneCabinet / NBS を利用（関連テキスト利用）

(3) 従来モデル

実験対象の番組映像として、1 本あたり約 20 分程度のニュース番組 16 本、約 90 分のサッカー中継番組 3 本、及び、ニュース原稿とサッカーのスコアシートといったテキストデータを用意した。番組毎の生成するメタデータの内容は表 5.4.1-1 の通りである。

表 5.4.1-1： 生成するメタデータの内容

	区間メタデータ	意味メタデータ
ニュース	ニューストピックの開始時間、終了時間	ニュースタイトル、概要、キーワード
サッカー	シュートシーンの開始時間、終了時間	タイトルのみ（シュートを打った選手名、チーム名）

作業者による作業時間の個人差や番組内容による変動を吸収するため、各作業者とも各作業モデルについて、複数回の作業を実施し、作業時間の平均値等を比較した。なお、各作業モデルとも入力映像に対し、「メタデータ自動生成エンジン部」におけるメディア解析処理は実サービス運用上では夜間にバッチ処理させる等で実施することが想定できることから、本実験では人手作業の時間効率化に関する評価、考察を目的とし、メディア解析処理後の人手作業にかかる時間を計測対象とした。

提案モデル A,B の具体的な作業内容としては、区間メタデータについては、代表テロップ当のメディア解析結果やニュース原稿、サッカーのスコアシート中の時間情報を手がかりに、SceneCabinet / NBS のキー画像ブラウザと再生モニタを利用して設定する。意味メタデータは、メタデータエディタ上の意味メタデータ生成ボタンを押下することで自動生成される情報を作業者が目視で確認し、間違いがあれば修正し、完成・保存する作業フローとした。意味メタデータの生成ルールを表 5.4.1-2 に示す。ニューストピックの意味メタデータのうち、キーワードの生成基準としては、10 個の単語を候補として一旦自動抽出した後、その中から、人名、時事用語、国名、市町村名、団体名を中心に作業者が主観で 5 個の単語を選択することとした。

表 5.4.1-2：意味メタデータの生成ルール

	ニュース			サッカー
	タイトル	概要	キーワード	タイトル
提案モデルA	代表テロップの認識結果を確認、修正	音声認識結果を確認、修正	タイトル、概要から単語抽出、選択	テロップ認識結果を確認、修正
提案モデルB	代表テロップの認識結果を確認、修正	読み原稿テキスト	タイトル、概要から単語抽出、選択	スコアシート中のテキスト
従来モデル	代表テロップの内容を書き起こし	アナウンサーの音声を書き起こし	タイトル、概要から単語を目視で抽出し、書き起こす	テロップ、音声などを確認し、書き起こす

なお、従来モデルの作業内容としては、区間メタデータの生成は、代表テロップは表示しないキー画像ブラウザと音声レベル表示機能の無い再生モニタを利用して実施し、意味メタデータについては、映像内容を目視で確認し、全て一から書き起こすこととした。以降では、ニュース番組、サッカー中継番組の番組ジャンル毎に実験結果を述べる。

5.4.2　実験結果

ニュース番組、サッカー中継番組の番組ジャンル毎に実験結果を述べる。

5.4.2.1　ニュース番組に対する実験結果

5.4.2.1.1　結果の概要

図 5.4.2.1.1-1 にニュース番組に対する実験結果のグラフを示す。作業モデル別に、4 名の作業者の作業時間の平均値を番組時間長で正規化して表現したものである。提案モデル A は映像時間の約 1.8 倍、提案モデル B は約 1.3 倍、従来モデルは約 3.6 倍であった。提案モデル B の作業時間は従来モデルを約 64%短縮したことになる。作業時間の内訳として、区間メタデータ、意味メタデータともに従来モデルの作業時間を最大で約 74%短縮できた。

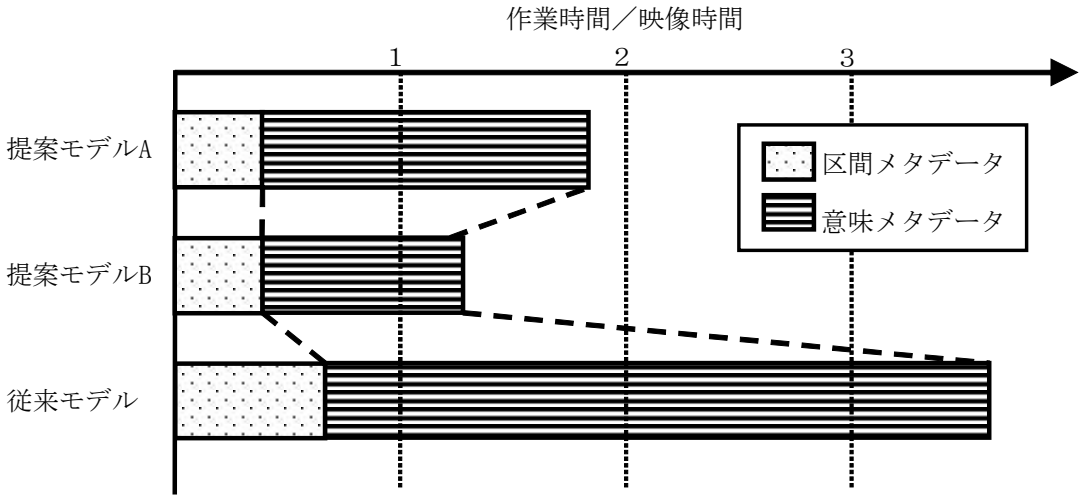


図 5.4.2.1.1-1：メタデータ生成作業時間の計測結果（ニュース番組）

また、生成されたメタデータのクオリティを表 5.4.2.1.1-1 に示す。区間メタデータ、意味メタデータともに全作業モデルにおいて再現率や文字一致率が100%に近い高品質のものが得られた。すなわち、SceneCabinet / NBS による提案モデルでも、特に台本テキストも合わせて利用できる場合（作業モデル B）においては、高品質のメタデータが生成でき、かつ、従来モデルに対し、顕著な作業時間の短縮効果があることが確認された。以降では、区間メタデータ、意味メタデータの各生成作業モデルについての実験結果を作業時間、メタデータのクオリティ、及び、作業者からのアンケート結果の各観点から詳細に述べる。

表 5.4.2.1.1-1　ニュース番組のメタデータのクオリティ

	区間メタデータ	意味メタデータ		
		タイトル	概要	キーワード
提案モデルA	再現率　100% 適合率　100% 時間差標準偏差　110.4msec （代表テロップ検出率　100%）	文字一致率　100% （代表テロップ文字認識率　96.5%）	文字一致率　98.7% （音声認識率　92.3%）	正解率　95.8%
提案モデルB	再現率　100% 適合率　100% 時間差標準偏差　97.6msec （代表テロップ検出率　100%）	文字一致率　100% （代表テロップ文字認識率　96.5%）	文字一致率　100% （読み原稿をそのまま利用）	正解率　85.1%
従来モデル	再現率　98.5% 適合率　100% 時間差標準偏差　103.4msec	文字一致率　100%	文字一致率　95.7%	正解率　96.3%

5.4.2.1.2　区間メタデータ生成の作業コスト評価

区間メタデータの生成作業について、まず、各作業モデルにおける作業要領を説明する。提案モデル A, B では、SceneCabinet / NBS のキー画像ブラウザに図 5.3.1-1 のような代表テロップの画像だけを表示することで、一画面でニュース番組中の全ニューストピックの区切りが把握でき、簡単に区間メタデータが設定できる。

これに対し、従来モデルでは、キー画像ブラウザを必要に応じてスクロール操作しつつ、多くのサムネイル画像の中からニューストピックの区切りを探す必要がある。このような作業要領の違いに起因すると考えられるが、作業時間は提案モデル A, B が従来モデルの作業時間を約 43%削減できた結果となった。

また、図 5.4.2.1.2-1 には、4 人の作業者の作業時間の平均値、最大値、最小値を示す。提案モデル A と提案モデル B は、平均値はほぼ同じだが、提案モデル B のほうが最大値と最小値の差、すなわち、作業時間の個人差が大きかった。提案モデル A は、ニュースの代表テロップの画像という、明確な基準があったことが、個人差が小さく済んだ理由と考えられる。提案モデル B は、代表テロップの画像の他、読み原稿中の時間情報に対応する画像も合わせてメディアブラウザに表示されることから、提案モデル A に比べ情報過多となり、作業時間の個人差が大きくなったものと考えられる。作業者からのアンケート結果によれば、SceneCabinet / NBS の再生モニタ上の各種ジャンプボタンの操作への慣れ具合の差が個人差に顕著に表れたものと考えられる。再生モニタのボタンの使い方に、より詳細なルールを設ける等により、作業時間の個人差、また、平均作業時間も減ると考えられる。

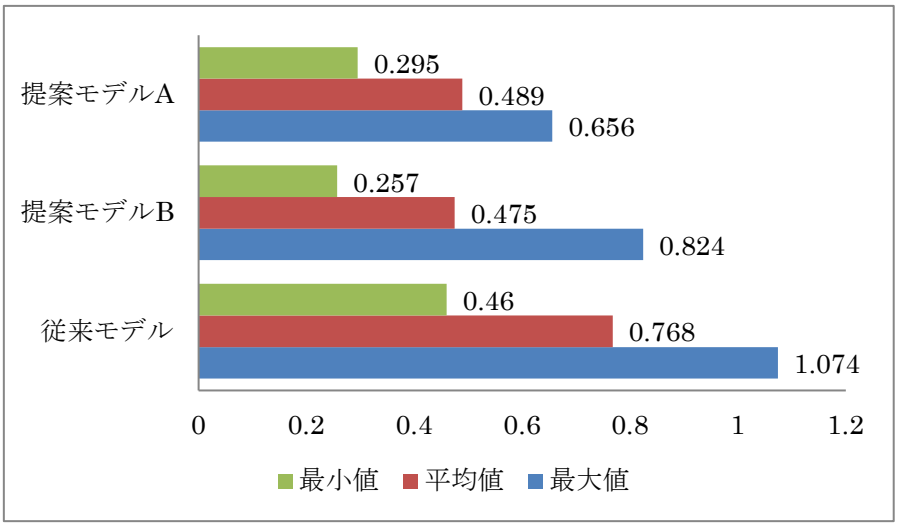


図 5.4.2.1.2-1　区間メタデータの生成作業時間（作業時間／映像時間）

また、生成された区間メタデータのクオリティに関しては、作業モデル A, B については、共に代表テロップの検出率が 100%であったことから、番組中の個々のニューストピック区間の抽出再現率は 100%であったのに対し、従来モデルでは 98.5%であった（適合率は全モデルとも 100%）。

従来モデルにおいては、キー画像ブラウザ上に全てのニューストピックの代表テロップが含まれる状態であったが、その他の多くのサムネイル画像の中から代表テロップ画像を探し出す際に見落としなどのミスが発生したため、ニューストピック区間の抽出再現率が 100%とはならなかった。

サービス要件にもよるが、ニューストピック区間の抽出再現率が 100%にならないことは作業品質として致命的であることが想定されることから、区間メタデータのクオリティの面では、代表テロップを利用する提案モデル A, B が従来モデルよりも非常に有効であるといえる。なお、事前に作成した正解データ（各ニューストピックの開始時間と終了時間）と生成した時間データとの差分について、そのばらつきを「時間差標準偏差」として定義し、表 5.4.2.1.2-1 に示した。

時間差標準偏差は、全ての作業モデルにおいて共通で、約 100msec 程度であり、他のニューストピックのシーン内容が含まれる等のサービス上問題がある程度の品質ではなく、良好な結果であったといえる。ニューストピックは、シーンの切り替わり等の明確な区切りが分かりやすい番組構成であるためと考えられる。

また、提案モデル A と提案モデル B の間においては、作業時間、クオリティとも有意な差は見られなかったが、作業者からのアンケート結果には、芸案モデル B において、台本テキスト中に含まれるニューストピックの放送予定時刻に対応するサムネイル画像はあまり有効ではなかったと意見があった。これは放送予定時刻と実際に放送されるニューストピックの放送時刻に数秒のズレがあるケースがあることが原因と考えられる。これより区間メタデータの生成には、台本テキストが利用できる場合においても、代表テロップの情報を優先的に利用することが作業時間の短縮に有効であると考えられる。

5.4.2.1.3　意味メタデータ生成の作業コスト評価

意味メタデータの生成については、テロップ認識、音声認識、及び、自然言語処理により自動生成されるニュースタイトル、概要文、キーワードがほぼ適切なものとして取得でき、一部を修正するだけで済んだことから従来モデルに比べ、提案モデル A, B の作業時間が短く済んだ（図 5.4.2.1.1-1）。特に、提案モデル B は従来モデルの作業時間を約 69%削減できた結果が得られた。

図 5.4.2.1.3-1 には、意味メタデータについて、4 人の作業者の平均作業時間に加え、個人レベルの最大値と最小値を示す。全ての作業モデルを通して、個人差が著しく大きいものはなく、最大値、最小値、平均値ともに、提案モデル B が最短、次いで、提案モデル A、従来モデルが最長であった。作業者からのアンケート結果によれば、特に、SceneCabinet / NBS のメタデータエディタで、概要文を確認・修正する操作の際に、個人差がより顕著に出たものと考えられる。作業者によっては、音声認識結果を修正する際に、映像内容を確認する操作を入念に複数回行うケースもあり、この操作時間の個人差の影響が強いと考えられる。音声認識結果の修正方法等、メタデータエディタの使い方に対し、より詳細なルールを設ける等により、作業時間の個人差、また、平均作業時間も減るものと考えられる。

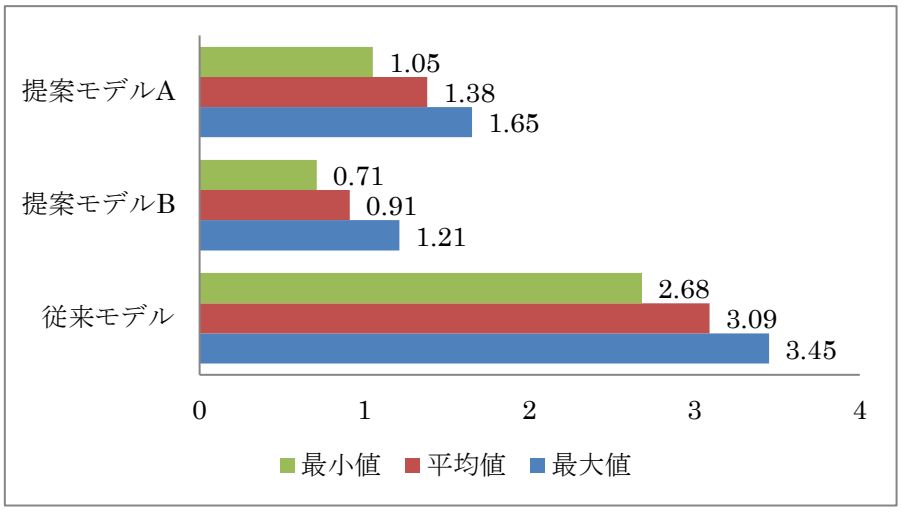


図 5.4.2.1.3-1　意味メタデータの生成作業時間（作業時間／映像時間）

以上は意味メタデータ全体の生成作業時間の計測結果と分析であるが、意味メタデータの構成要素である、ニュースタイトル、概要文、キーワードのクオリティや修正作業の手間にはそれぞれ異なる特徴がみられた。

ニュースタイトルは、作業モデル A, B と従来モデルとの間に作業時間の有意

な差は見られなかった。作業モデル A, B においては、代表テロップの文字認識精度は 95%以上と良好であったが、実際に認識結果に間違いがないかを結局映像を見て確認する必要があった。この作業と従来モデルにおいて、映像中のテロップを目視で確認し、一から手入力する作業は、映像内容を確認する操作が作業の大部分を占める点で、実質さほど変わらない作業内容であったといえる。

概要文については、提案モデル B がニュース原稿のテキストをそのまま修正なしで利用できることから、提案モデル A や従来モデルよりも格段に短い作業時間で作成できた。提案モデル A は、ニュースタイトルと同様、音声認識結果を確認、修正する作業が従来モデルと実質同程度の作業時間であった。提案モデル A における音声認識結果の精度（文字一致率）は、台本テキストを正解とした場合、92.3%であった。これを実際の音声を聞きながら修正した結果が 98.7%と 100%にはならなかった。

従来モデルも同様であるが、音声を聞いてテキストを書き起こした結果、実際の番組でのアナウンサー音声の内容と正解データである台本テキストの内容が文章の終わりの言い回しなどの細かい部分で異なったり、作業者によって、一部、タイピングの仮名漢字変換を誤ったりするケースがあったことが原因と考えられる。ただし、文章中の重要な固有名詞や名詞にはほとんど間違いはなく、サービス利用にも十分耐えうる品質が得られていると考えられる。なお、概要文はニュースタイトルに比べ、文字数が多いため、音声認識結果の誤り箇所だけを修正するだけで済む提案モデル A のほうが、全ての文字をタイピングする従来モデルに比べ、作業の疲労感は少なかったという感想が得られた。

キーワードについても、提案モデル B が最も短い作業時間であった。提案モデル B では、ニュース原稿のテキストから、参考文献[hayashi03]の方式を利用して自動生成される 10 個の単語の中に人名、時事用語などの単語が正しく含まれたため、主観で 5 個を選択する作業も容易に実施できた。

しかしながら、最新の時事用語など、参考文献[hayashi03]の方式に必要な単語辞書に含まれない用語については、10 個の候補の中に上がっていないこともあり、最終的に 5 個に絞り込んだキーワードが別途、事前に同様の生成基準で作成しておいた正解キーワードに含まれる割合（正解率）は 85.1%であった。

提案モデル A と従来モデルについては、ニュースタイトル、概要文と同様、映像内容を確認する作業が入るため、作業時間としては、提案モデル B よりも長くかかった。ただし、最新の時事用語なども漏れなくキーワード化できたため、正解率は 95%以上と提案モデル B よりも高いものであった。実運用時には、提案モデル B において、単語辞書の最新化を高頻度に実施することで、クオリティの高いキーワードを短時間に生成することが可能になると考えられる。

5.4.2.2　サッカー中継番組に対する実験結果

5.4.2.2.1　結果の概要

図 5.4.2.2.1-1 にサッカー中継番組のメタデータ生成作業時間の計測結果を示す。区間メタデータの生成作業は、表 5.4.2.2.1-1 のとおり、番組中のシュートシーンの開始時間と終了時間を設定する作業である。作業時間としては、メディア解析結果とスコアシートの両方が利用できる提案モデル B が最も短時間であり、最も時間がかかった従来モデルの作業時間を約 59%削減できた結果が得られた。区間メタデータのクオリティに関しては、三つの作業モデル間で異なる傾向が見られた。クオリティを評価する際の正解データとして、番組中のシュートシーンの個数は、スコアシート中に含まれるシュート情報の個数を利用し、各シュートシーンの開始時間と終了時間は実際に映像内容を確認し、シュートが打たれた瞬間のタイミングの 10 秒前を開始時間、10 秒後を終了時間とする規定で作成した。以降では、区間メタデータ、意味メタデータの各生成作業モデルについての実験結果を作業時間、メタデータのクオリティ、及び、作業者からのアンケート結果の各観点から詳細に述べる。

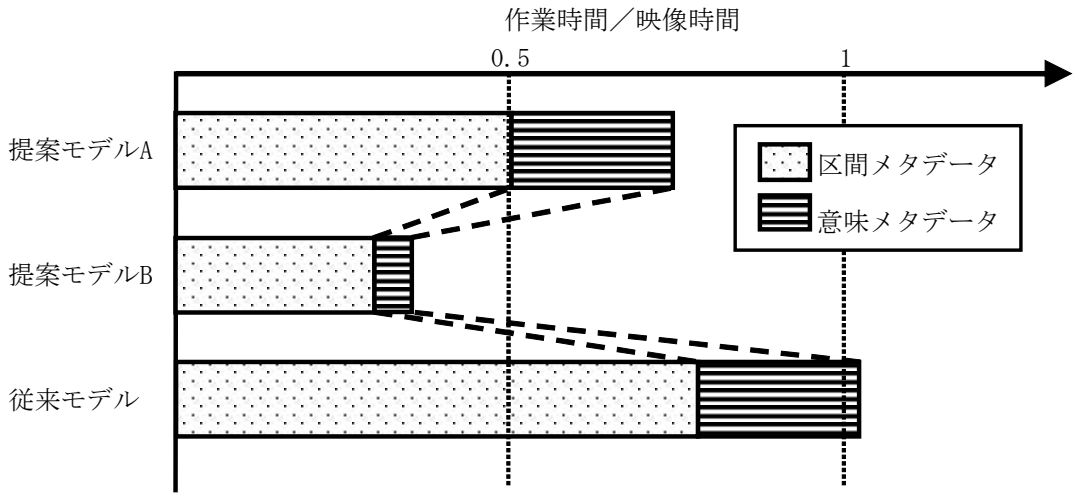


図 5.4.2.2.1-1：メタデータ生成作業時間の計測結果（サッカー中継番組）

表 5.4.2.2.1-1　サッカー番組のメタデータのクオリティ

	区間メタデータ	意味メタデータ
		選手名とチーム名
提案モデルA	再現率　80.6% 適合率　100% 時間差標準偏差　2.3sec （代表テロップ検出率　80.6%）	正解率　100%
提案モデルB	再現率　100% 適合率　100% 時間差標準偏差　2.5sec （代表テロップ検出率　80.6%）	正解率　100%
従来モデル	再現率　77.5% 適合率　100% 時間差標準偏差　2.7sec	正解率　100%

5.4.2.2.2　区間メタデータ生成の作業コスト評価

提案モデル A においては、サッカー番組において、シュートシーンの後にはシュートを打った選手の名前のテロップ文字が表示されるという特徴に着目し、この選手名のテロップ文字を代表テロップとして検出することで、シュートシーンを探すための手がかりとした。すなわち、キー画像ブラウザ上の代表テロップのサムネイル画像を選択し、その時間から数十秒程度遡った時点の映像内容を確認するという作業要領である。

今回の実験で用いたサッカー番組におけるテロップ文字表示の映像特徴については、選手名のテロップ文字の文字色は番組中を通じて同一色であったが、文字に接する背景色が文字の表示回毎に変わり、また、文字の表示時間については、瞬間的に表示され、すぐに消えてしまうものがある等、ニュース番組のタイトルテロップほど文字色と背景色のコントラスト、及び、文字の表示継続時間が一定ではなかった。このため、画像中の輝度エッジ情報の時空間的に密集度合いを評価する[C2]の方式を用いても、ニュース番組のタイトルテロップほど高い検出率を得ることができなかった。

このため、メディア解析処理による代表テロップの検出漏れがあり、最終的に生成したメタデータ中のシュートシーンの再現率は約 80%であった。また、番組中にごく稀にシュートシーンの後に選手名のテロップが表示されないケースもあった。このため、本実験としては、一定以上のシュートシーンの抽出再現率を得ることはできたが、例えば、実用上ですべてのシュートシーンを確認するような要件が求められる場合においては、提案モデル A のアプローチは適切ではないことになる。

提案モデル B では、スコアシート中に記録されるシュートシーンの時間情報（図 5.3.3-1 (b)）を手がかりに、代表テロップ等、その時間付近のサムネイル画像だけキー画像ブラウザ上でチェックする作業で済んだため、提案モデル A よりも短時間であったとともに、クオリティについても前記の正解データの定義より、再現率は 100%であった。ただし、スコアシート上のシュートの時間情報には、数十秒レベルで誤差があることもあり、キー画像ブラウザ上で、スコアシートの時間情報を手がかりに、その時間付近のサムネイル画像、特に選手名のテロップ文字を手がかりとして、区間メタデータを設定することが作業時間の短時間化には有効であったといえる。

先に述べた提案モデル A では、スコアシートを用いないため、シュートシーン以外において表示される選手名のテロップを代表テロップとして誤検出するケースに対し、これが本当にシュートシーンの後の選手名のテロップ文字かどうかチェックする必要のある作業内容であった。これに対し、提案モデル B では、最初から正解であるスコアシート中のシュートシーンの時間を手がかりとするため、本当にシュートシーンかどうかをチェックしなくともよく、提案モデル A に比べ、作業者の精神的な負担も少ないというアンケート結果も得られた。

また、従来モデルでは、提案モデル A よりも数倍の量のサムネイル画像の中から選手名のテロップを中心にシュートシーンを探す作業内容であるが、キー画像ブラウザ上での情報の一覧性が悪く、提案モデル A、B の約 2 倍の作業時間であった。メタデータのクオリティについても、作業中に選手名のテロップ文字を見落とすこともあり、シュートシーンの抽出再現率は約 78%であった。

このことから、キー画像ブラウザに表示するサムネイル画像の量や、その中に含まれる有用な情報の割合が、作業時間やメタデータのクオリティに大きく影響するものと考えられる。時間差標準偏差については、全モデルにおいて、約 2.5 秒とニューストピックのメタデータよりも値が増えた。シュートが打たれる瞬間のタイミングが、区間メタデータ生成の基準となるが、シュートが打たれるタイミングの判断に個人差が現れたことが原因と考えられる。

図 5.4.2.2.2-1 には、4 人の作業者の作業時間の平均値、最大値、最小値を示した。メディア解析結果とスコアシートの情報を両方活用する提案モデル B が提案モデル A に比べ、平均作業時間が最も短い他、最大値と最小値の差、すなわち、作業時間の個人差も少ない結果となった。提案モデル B では、特に、作業者全員が、区間メタデータの作り方の要領に慣れたものと考えられる。一方で、従来モデルは平均作業時間が最も長い他、個人差もかなり大きいものとなった。シュートシーンを見逃さないように、ずっと再生モニタの画面を集中して見続ける必要があり、作業の途中で疲労具合により、作業時間に個人差が出るものと考えられる。また、全モデル共通して、サッカーのルールに詳しい作業者と詳しくない作業者との間でも、作業時間に個人差が生じた。サッカーのようなスポーツ番組へのメタデータ生成には、そのスポーツのルール等にどれだけ精通しているかも作業コスト削減には重要な要素になると考えられる。

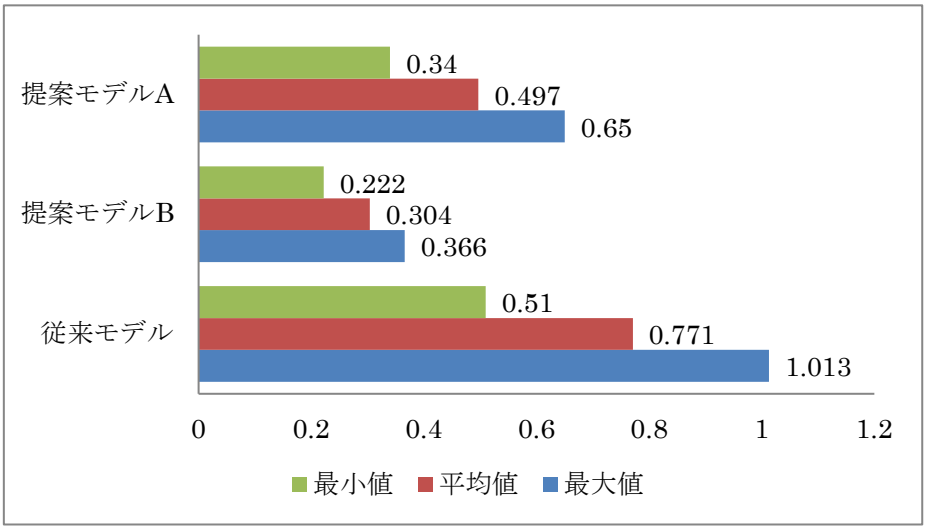


図 5.4.2.2.2-1　区間メタデータの生成作業時間（作業時間／映像時間）

サッカー番組は、基本的にはサッカーのプレー映像が継続的に続く映像内容であり、ニュース番組ほど番組構成が明確ではないことから、時間設定の作業には個人差が生まれやすいものと考えられる。また、全モデルを通じてシュートシーンの抽出適合率は 100%であった。例えば、提案モデル A では、メディア解析時に代表テロップの誤検出があったが、作業者がこれを誤ってシュートシーンとして定義することはなく、正しく判断できたということになる。

5.4.2.2.3　意味メタデータ生成の作業コスト評価

意味メタデータは、シュートシーンにおいて、シュートを打った選手名とチーム名である。図 5.4.2.2.1-1 に示した通り、提案モデル B では、スコアシート中の選手名、チーム名の情報をそのまま流用できることから、ほとんど作業時間はかからなかった。提案モデル A においては、サッカー中継番組の場合、選手名、チーム名テキスト情報のテロップ文字表示、実況音声ともにニュースに比べ、参考文献[C2][Ohtsuki]の技術での自動認識が困難であったことから実質、一から手作業で入力する従来モデルと同じ作業内容となった。文字認識については、文字部分の画像解像度が非常に低く、また音声認識については、歓声音などの実況以外の音が含まれる等が自動認識困難の主な理由である。このため、提案モデル A と従来モデルはほぼ同じ平均作業時間となった。

図 5.4.2.2.3-1 には、意味メタデータについて、4 人の作業者の平均値、最大値、最小値を示す。提案モデル B は、上述の通り、スコアシートの情報をそのまま利用できることから、単純な作業内容で済み、最大値と最小値の差、すなわち、作業時間の個人差は小さいものとなった。一方で、提案モデル A は、上述の通り、メディア解析が困難であったことから、従来モデルと平均値がほぼ同じとなったことに加え、最大値と最小値の差もほぼ同じとなった。提案モデル A はメディア解析の結果が実質活用できず、作業内容がほぼ従来モデルと同様となったことが原因と考えられる。今後、シュートシーンの特徴を捉えたメディア解析技術の検討が重要と考えられる。

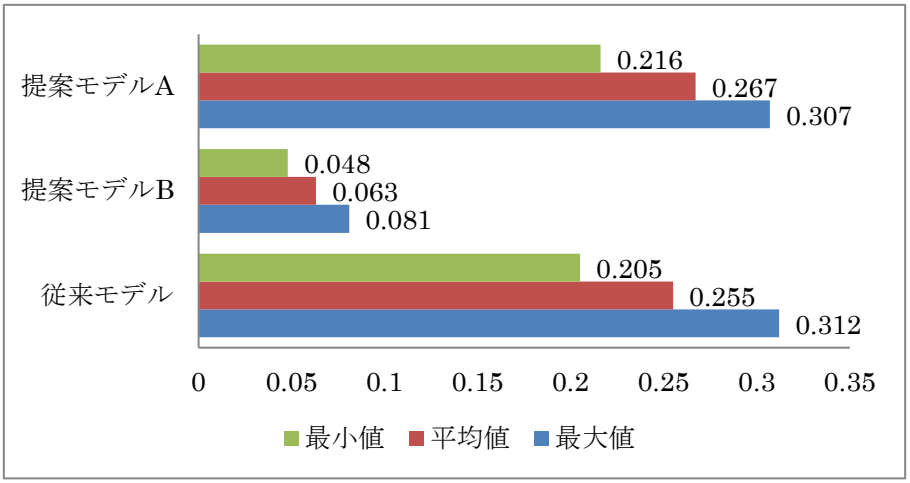


図 5.4.2.2.3-1　意味メタデータの生成作業時間（作業時間／映像時間）

以上のことから、サッカー番組のシュートシーンのメタデータ生成には、スコアシートの情報も利用する提案モデル B が作業時間、メタデータのクオリティの観点から優良な作業モデルであるといえる。ただし、区間メタデータの生成については、スコアシートの情報だけではなく、これと映像解析処理の結果を組み合わせることで作業時間の短縮化に最大の効果を発揮することが分かった。

5.4.3 考察

以上の結果より、代表テロップの検出を中心とするメディア解析処理、ユーザインタフェース、及び、台本テキストの利用が特徴である提案作業モデルがもたらすメタデータ生成の効率化効果について考察する。

メディア解析技術については、特に、ニュース番組の場合、代表テロップの検出・認識、音声認識、自然言語処理により、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ることが、作業時間の削減に大きく寄与すると考えられる。代表テロップの検出は従来システムにはない新機能であり、キー画像ブラウザにおいてニューストピックの区切りやシュートシーンを見つけるのに非常に有用であった。ただし、サッカー中継番組においては、シュートシーン以外の代表テロップ情報も多数抽出されたことから、ニュース番組に比べ、最終的に生成したいメタデータとはまだ乖離があったといえる。

しかしながら、サッカー中継番組からシュートシーンだけを正確に自動抽出できる汎用的なメディア解析方式はまだ確立されていない。サッカー中継番組だけでなく、さまざまな映像ジャンルから所望のメタデータが自動取得できるよう、メディア解析処理の性能を一層向上させていくことが、作業コスト削減に向けた今後の重用な課題となる。

また、今回提案した代表テロップを用いた区間メタデータの生成方法は、番組ジャンルに非依存で一定以上の効果を発揮できるレベルではなく、利用するに適した範囲で効果的な使い方を検討するものである。今回、代表テロップの検出には[C2]で提案している方法を用いたが、テロップ文字周辺の輝度エッジ情報から文字列矩形としての幅と高さを数ピクセル内の誤差精度で求め、これと予め設定した代表テロップとしての矩形条件とを比較する手法である。

このため、なんらかの方法で事前に代表テロップとしての条件を取得しておく必要がある。このことは、実際のテレビ番組に対して行うことを想定すると、

1 回しか放送しないような番組については、代表テロップの条件を調べても、調べる行為そのものが代表テロップの検出作業になっており、[C2]の手法を用いて再度検出する必要が必ずしもあるわけではない。

また、ニュース番組のニュースタイトルテロップのように、他のテロップとの文字列矩形の幅と高さの違いの度合いが高いテロップが使われており、かつ、そのテロップ表示のタイミングが意味的なシーンの区切りのタイミングと高い相関がある場合に有効な手法である。すなわち、今回の実験で用いたニュース番組のニュースタイトルやスポーツ番組の選手名のテロップの他、情報・バラエティ番組の各トピックのタイトルテロップ、あるいは、音楽番組の曲名やアーティスト名のテロップ等、1 回限りの放送ではなく、ある程度の期間で複数回に渡り放送される番組であり、その間、同一条件でテロップ表示が施される番組に対して適用することで、意味のあるシーン区間を抽出するのに有効なものである。

この条件を満たす番組であり、長期間放送が続くような番組に対しては、最初に代表テロップの条件を調べておくだけで、あとは特別な作業無しに放送期間中は自動的に有用なメタデータの手がかりが得られるということになる。

また、ユーザインタフェースのデザイン効果については、区間メタデータの生成作業においては、作業者からのアンケート結果より、キー画像ブラウザ上部のイベント種別選択ボタンによるシーンイベント単位でのサムネイル画像表示の切り替え機能、及びイベント毎にサムネイル画像を色分け表示できる点が作業時間短縮化に寄与できたといえる。例えば、代表テロップのサムネイル画像だけをキー画像ブラウザ上に表示することが簡単な操作ででき、これがメタデータ生成作業の短時間化に大きく寄与したものと考えられる。

また、同じくアンケート結果から、再生モニタ上の表示映像の時間位置を微調整する作業においては、再生モニタ上の±10 秒、±1 秒といった各種ジャンプボタン、音声レベルの波形表示を利用することで、作業中の映像再生、早送りの時間を減らし、作業の短時間化に有効であったという GUI 効果に関する意見も得られた。

キー画像ブラウザ上のサムネイル画像と再生モニタは連動し、サムネイル画像のシーンからのジャンプ再生表示が行えるが、再生モニタ上の各種ジャンプボタン、音声レベル表示については、キー画像ブラウザ上の複数のサムネイル画像の時間の間に存在する所望のシーンを探す作業を短縮化するのに有効であったといえる。このようになるべく映像再生する時間を割く必要なく、所望のシーンが探し出せる GUI 設計が作業短時間化のためには重要であると考えられる。

台本テキストの有効性については、ニュース番組の区間メタデータ生成の際のように、メディア解析結果のほうが信頼できるケース、逆に、意味メタデータ生成の際のように台本テキストがそのままメタデータとして再活用でき、作業短時間化に大きな効果をもたらすケースがある。当然ではあるが、台本テキストの効果は、その内容と生成したいメタデータとの違いに依存するものでる。

区間メタデータ生成への効果は、台本テキスト中の時間情報がどのようなシチュエーションで作成されたかに依存すると考えられる。ニュースの場合は、放送前に正確な放送予定時間として作成されるが、サッカーのようなスポーツ番組においては、試合中にリアルタイムに記録されるため、実際の映像の時間とは 1 分程度異なる場合もあり、かなりラフな情報である。台本テキストの内容に応じた、的確な利用方法を選択していくことが作業コスト削減には重要である。

また、作業手順の明確さも作業コストに大きく影響すると考えられる。例えば、今回の実験では、作業内容の指針をある程度明確にルール化できたため、作業者に特に専門のスキルがなくても、作業中にさほど迷うことなく作業を円滑に進めることができた。

ただし、作業手順の内容が少しでも変わると、作業時間が大きく変わることも考えられる。例えば、今回の実験では、ニューストピックのキーワードの個数として「5 個」としたが、これを「1 個以上」という緩い規定にした場合、作業者によっては、ニュースタイトル中の代表的な単語 1 個だけを選択して簡単に済ませてしまうケースや、逆に、映像内容をじっくり把握して、複数の単語をじっくり吟味したうえで設定したり、また、どの単語を選んで良いか迷ったりするケースも考えられ、作業者によって作業時間やメタデータのクオリティの統一性が保ちにくい状態になる。すなわち、作業手順の内容に応じて、作業時間、メタデータのクオリティは変わる要素がある。今回の実験結果については、ある作業モデルを設定した場合の結果例としてとらえておき、将来的に実サービスを想定した場合の基礎データという位置づけとなるものと考えられる。

また、作業者に専門スキルも必要なケースがある。サッカー等のスポーツ番組の場合のスポーツルールへの精通度のようなスキルもあるが、他にも、生成したいメタデータを映像視聴サービスでどのように利用していくかという点においても専門性が必要なケースがある。例えば、番組全体のダイジェスト映像をストーリー立てて配信するようなサービスの場合、ドラマ性のあるダイジェスト映像としての演出効果も含んだメタデータを生成するというクリエイティブなスキルが必要になる。

ただし、このようなケースにおいても、ダイジェストを作成するために必要と考えられるシーンを全て素材シーンとして一旦事前に作成しておき、その後、素材シーンを基にクリエイターがメタデータを完成させるという 2 段階に分けることができれば、前半の素材シーンの作成作業は、作業ルールが定義しやすく、かつ、クリエイターの作業も減り、作業全体のコスト削減が見込めることが想定される。作業全体の中から作業内容の指針が明確にルール化できると作業が切り出せることも作業コスト削減には重要なポイントとなると考えられる。

また、ニュース番組、サッカー中継とも、区間メタデータ生成、意味メタデータ生成の作業時間に一定の個人差もあることが確認された。両番組ともに、区間メタデータ生成作業時間の個人差が大きく、意味メタデータ生成の作業時間、特に、サッカー中継のシュートシーンの意味メタデータ生成において、スコアシートを利用するケースは、区間メタデータ生成に比べ、個人差がない結果となった。3.1.2 節で述べたが、メタデータ生成作業を複数人で分業できる可能性については、この個人差を認識した上で、分業方式を検討する必要がある。本章における実験結果を踏まえると、区間メタデータ生成、意味メタデータ生成、共に、作業内容、ルールを極力シンプル、かつ、分かり易くしていくことで、作業時間の個人さが減り、複数人での分業スタイルの効果がみられる可能性はあるものと考えられる。例えば、個人差が大きい、区間メタデータの生成は、慣れた専門の担当者が担い、個人差が少なかった意味メタデータの生成は、アルバイトの担当者が担う等の運用が考えられる。

5.5 まとめ

本章では、制作済番組を対象とし、メタデータ生成の作業コスト削減に関する作業モデルを提案し、実験結果と考察を述べた。作業モデルを実現するメタデータ生成の作業者向けのユーザインタフェースシステムとして「SceneCabinet / NBS」を提案、実装した。SceneCabinet / NBS はメタデータ生成の作業において、特に、従来、人手作業に時間がかかっていた区間メタデータの生成作業を短縮化する様々な機能を含め、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 部から構成されるものである。

SceneCabinet / NBS を用いた、区間メタデータ、意味メタデータの具体的な生成作業内容を提案した。テロップ文字の大きさ等の視覚的特徴に加え、カット点、人物動作、音声認識、そして、台本テキスト等の情報を組み合わせ、ニュースストピックや野球中継の各打者のシーンに関する区間メタデータ、意味メ

タデータを生成する作業モデルを提示した。実験により、提案作業モデルの作業コスト効果を確認し、メディア解析、ユーザインタフェース、台本テキストの有効性を明らかにする考察を行った。

実験により、ニュース番組全体から個々のニューストピックに関する区間メタデータ、意味メタデータを生成する作業において、SceneCabinet / NBS を用いる作業モデルは、全て人手作業で行うモデルの作業時間を約 64%短縮できることを明らかにした。これには、区間メタデータが 100%自動生成できたことが大きく寄与したもの考えられる。意味メタデータも 69%という大幅な時間短縮ができた。自然言語処理により、タイトル、概要文、キーワードの候補テキストが自動で得られることが寄与したものと考えられる。

また、生成されるメタデータの精度の観点においては、ニューストピックの区間メタデータや意味メタデータはほぼ正解と同様の内容だったが、サッカー中継番組のシュートシーンについては、特に、シーンの開始時間は作業者間で 2, 3 秒程度の違いが生じた。メタデータ生成の作業ルールの明確さや作業者の映像内容に対する背景知識等が大きく影響することを明らかにした。これらは、今後のメタデータ生成の作業ルール策定に大きく寄与できる情報になると考えられる。

第6章 ライブ番組向けメタデータ生成の作業コスト削減

6.1 概要

前章では制作済の番組を対象としたメタデータ生成の作業コスト削減に向けた作業モデルの提案、実験と考察を述べた。本章では、メタデータを利用した番組視聴サービスの適用範囲の拡大を指向し、ライブ放送される番組を対象として、番組放送中にリアルタイムにメタデータを生成する技術、作業モデルを提案し、人手作業との比較実験の結果と考察を述べる。

ライブ番組のメタデータ生成は、番組の放送中に、番組進行に遅れることなく、区間メタデータと意味メタデータを次々に生成するものである。我々は、前章で述べた、メタデータ生成用のユーザインタフェースシステム SceneCabinet / NBS をライブ放送用に拡張した「SceneCabinet / Live!」を開発した。

SceneCabinet / Live! は、メタデータ生成の作業者に対し、キーボードでのタイプ入力無しに、音声で情報入力するだけで、区間メタデータ、意味メタデータが生成できる、より簡便ないユーザインタフェースを提供する。作業者の発話内容にルールを設けることで、誤りのない音声認識結果を得て、そのままメタデータとして活用するメタデータ生成の作業モデルを提案する。

以降、6.2 節では、ライブ番組のメタデータを活用した視聴サービスとメタデータ生成作業に求められる要件を述べる。6.3 節では、ライブ番組向けのメタデータ生成の作業モデルとユーザインタフェースシステム SceneCabinet / Live! を提案する。6.4 節で、ライブ野球中継に対し、提案作業モデルによるリアルタイムでのメタデータ生成実験と結果、及び、考察を述べる。6.5 節で、本章のまとめを述べる。

本章の内容は、筆者らの公表論文[J2],[C5]に基づくものである。

6.2　ライブ番組向けメタデータ生成

6.2.1　ライブ番組のメタデータサービス

図 6.2.1-1 に、野球中継を例として、ライブ番組のメタデータを活用したダイジェスト視聴サービスの画面例を示す。テレビ画面には、番組の進行に合わせて、ダイジェスト視聴用のシーンに関するメタデータが時々刻々増えて表示されていくイメージである（図 6.2.1-1 のテレビ画面の左下）。番組視聴者は見たいシーンに関する情報を選択すると、そのシーンのリプレイ映像が視聴できる、いわゆるタイムシフト視聴が番組進行中に実行できるものである。

野球中継での好きな選手のシーンの他、ニュース番組中の注目トピック等、ライブ番組中の注目シーンを見逃してしまった場合やもう 1 度見たい場合に、好きなタイミングで簡単に該当シーンをリプレイ視聴することが可能になり、テレビ視聴時間を有効に活用できる。このような新しい番組視聴サービスの実現にあたっては、番組放送中にリアルタイムに区間メタデータと意味メタデータを生成しなければならない。

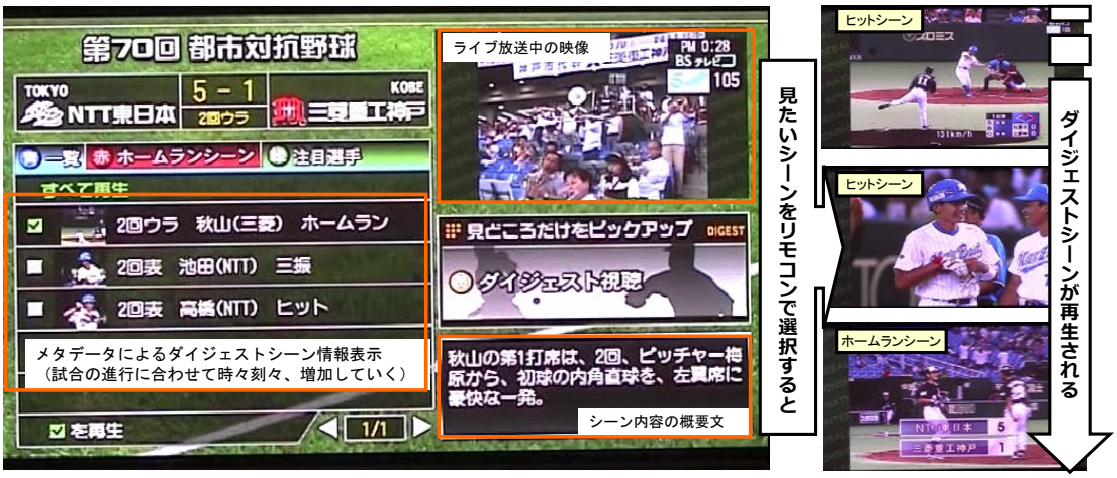


図 6.2.1-1　メタデータを利用したライブ野球中継のダイジェスト視聴サービスの画面例

6.2.2　ライブ番組に対するメタデータ生成の要件

ライブ野球中継番組のダイジェスト視聴に必要なメタデータ項目を図 6.2.2-1 に示す。番組中の個々のシーン区間の開始時間、終了時間、およびシーン内容に関するタイトル、概要文、キーワードがメタデータ項目になる。これらの項目は TV Anytime Forum においてもメタデータの国際標準仕様として策定されている。

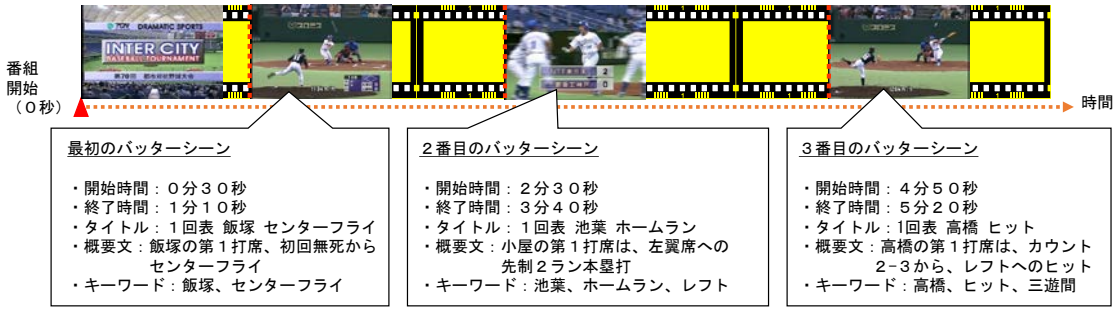


図 6.2.2-1　ライブ野球中継のダイジェスト視聴サービスに必要なメタデータの例

図 6.2.1-1 に示した、ライブ番組のダイジェスト視聴サービスを実現するためには、図 6.2.2-1 に示したメタデータを番組進行中にリアルタイムに作り出し、即時に、テレビ、パソコン、スマートフォンといった視聴者のネット端末に向けて送り出す必要がある。

メタデータの生成作業をすべて手作業で行うと膨大な手間が必要となり、作業の疲労感も大きくなる。特に、シーン内容に関するタイトル、概要文、キーワード等のテキスト情報については、第 5 章で述べたような番組台本等の既存テキスト情報が存在すれば、それを有効活用して効率的にメタデータとして生成できるが、スポーツ番組のように事前に筋書きの読めない番組については、実際にシーンが終わった段階で、初めてシーンのタイトル等の表現内容を検討し、書き起こす必要がある。

この作業を手作業だけで進めていくと、疲労感が大きいだけでなく、作業時間に余裕がなくなり、情報の入力ミスも生じやすくなることが想定される。そのため、ライブ番組の進行に遅れることなく、かつ、作業者の疲労感も少なくすむメタデータ生成の作業モデルを確立することが重要である。

6.3　SceneCabinet / Live! を用いたメタデータ生成の作業モデル

6.3.1　作業モデルの概要

野球中継のライブ放送を対象として、メタデータを番組進行に遅れることなく高速に生成していくための作業モデルを策定する。策定にあたっては、メタディア解析技術、ユーザインタフェース技術を最大限活用し、手作業の分量が少なくすみ、かつ作業者は野球中継の状況からなるべく目を離さずにメタデータ生成の作業が効率的に進めていけることを意識した。この考え方に基づき、我々は特に音声認識技術を積極的に勝つようした作業モデルを策定した。図 6.3.1-1 に策定した作業モデルを示す。また策定した作業モデルを実現するためのユーザインタフェースシステムとして「SceneCabinet / Live!」を開発した。システムの画面例を図 6.3.1-2 に示す。

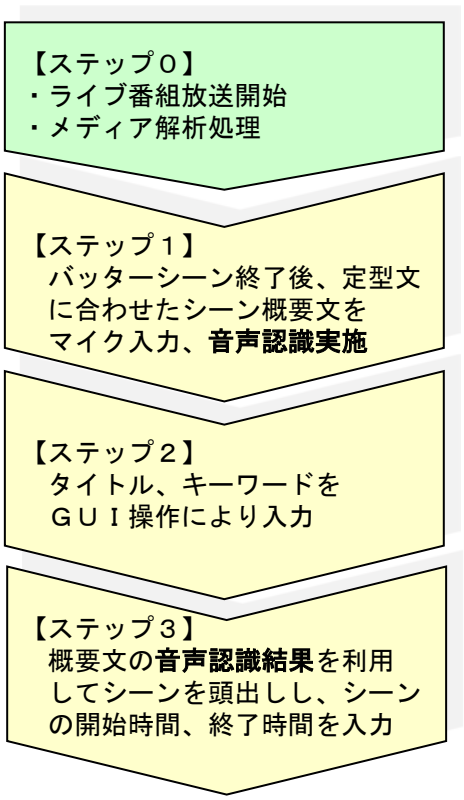


図 6.3.1-1　ライブ野球中継向けのメタデータ生成の作業モデル

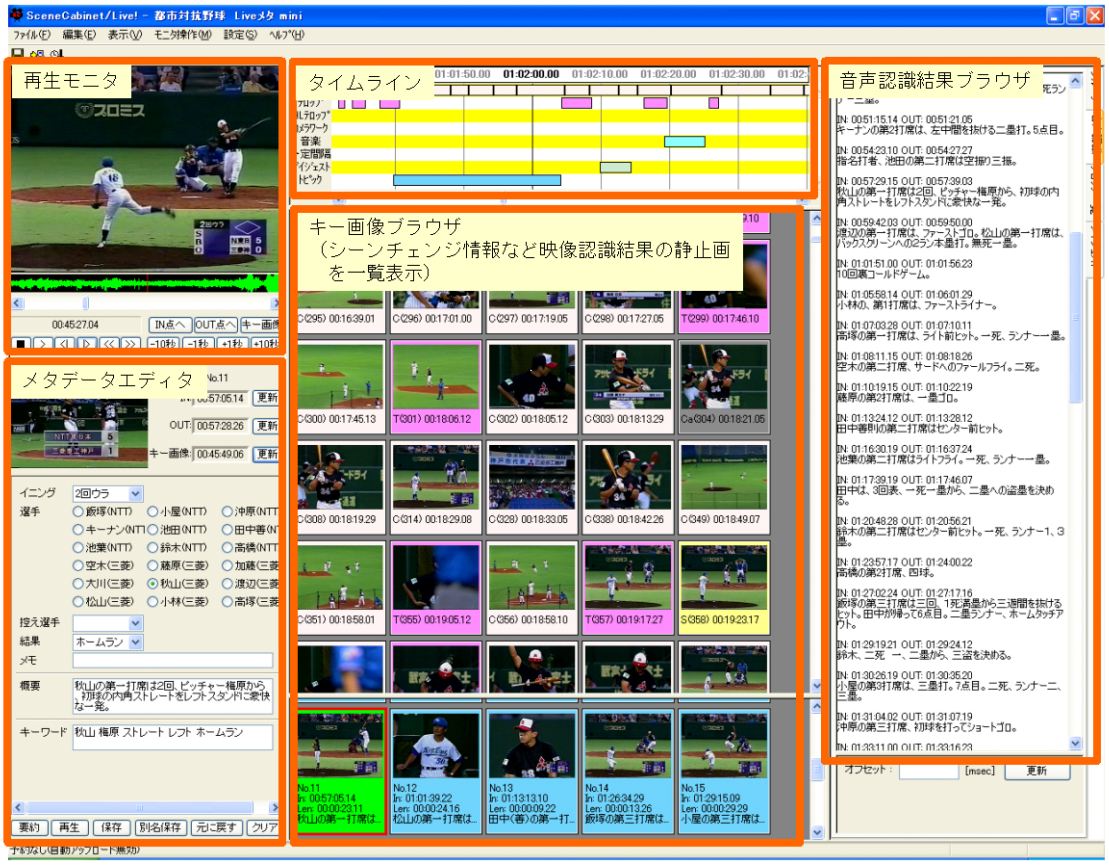


図 6.3.1-2　SceneCabinet / Live! の画面例

作業モデルとしては、まず、ホームランやヒットなどのシーン終了時点で、メタデータ生成の作業者がシーン内容を説明する概要文を音声認識しやすい条件下（周囲の雑音が少ない環境）で発話する。

概要文はシーン内容を説明する文章であり、テレビ画面の限られた表示スペースに表示される文字情報として利用されるため、文字数を考慮して設定する必要がある。そのため、定型フォーマットに合わせた文章として発話することとする。

図 6.3.1-3 に野球番組のバッターシーンに対する定型概要文の例を示す。定型文としての音声認識処理は、認識結果に対する制約条件を設けることができることから、自由発話の音声認識処理に比べ、高い認識率を得ることができ、シーン概要文の定型化が想定可能な番組のメタデータ生成に有効な方法である。

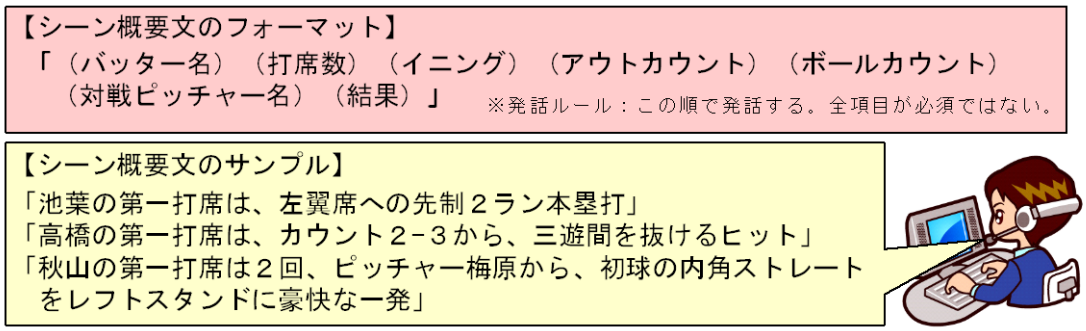


図 6.3.1-3　野球番組のバッターシーンに対する定型概要文の例

6.3.2　区間メタデータと意味メタデータの同時生成

実際の SceneCabinet / Live! の操作方法について述べる。メタデータ生成対象のシーンに対し、作業者が発話音声をマイク入力すると、音声認識結果された結果が図 6.3.1-2 の SceneCabinet / Live! の右側の音声認識結果ブラウザ上に表示される。作業者は、表示された音声認識結果ブラウザ上のテキストを、必要に応じて修正し、図 6.3.1-2 の SceneCabinet / Live! の左下のメタデータエディタに設定することで、シーンの概要文を完成させる。

シーン内容のタイトル、キーワードもメタデータエディタ上で入力する。タイトルは選手名やホームラン、ヒットなどの打席結果をラジオボタンなどの選択式 GUI を使って入力する。キーワードに関しては、メタデータエディタ上の要約ボタンを押すと、自然言語処理が実行され、タイトルと概要文としてすでに設定されているテキストから自動的に抽出される重要単語が得られ、これがキーワードとして設定される。

以上は、意味メタデータの生成方法だが、作業者の発話タイミングを工夫することで、同時に区間メタデータの生成も可能となる。すなわち、前述の仕組みに対し、作業者が概要文を発話する際に、そのタイミングをメタデータ入力対象のシーンの終了直後、例えば、打者がヒットを打ち、1 塁ベースまで進んだ後のタイミングで実施する。すると、SceneCabinet / Live! のシステム上は、音声認識の結果のテキストと共に、区間メタデータの終了時間も同時に記録される。

SceneCabinet / Live! の音声認識結果ブラウザには、区間メタデータの終了時間も表示されるため、このブラウザと、SceneCabinet / Live! の画面中央の

キー画像ブラウザ、および画面左上の再生モニタを組み合わせて利用することで、区間メタデータが少ない操作で確認できる。

具体的には、音声認識結果ブラウザ上の音声認識結果のテキストやキー画像ブラウザ上に表示される映像認識結果の代表画像を選択すると、選択されたそれぞれに対応する映像中のシーン画像が再生モニタ上に表示される機能を活用する。

すなわち、音声認識結果ブラウザ上のテキストを選択すると再生モニタ上では、最初から対象シーンの区間メタデータの終了時間の頭出しができる。その後、キー画像ブラウザ上の代表画像や再生モニタ上の±10 秒、±1 秒などのボタンで時間を微調整し、区間メタデータの開始時間を設定する。

以上の SceneCabinet / Live! の操作によれば、ライブ番組のダイジェスト視聴用のメタデータ生成作業に対し、テキスト情報の入力を 1 からキーボードなどをタイプしていく必要はなく、野球中継の状況を見ながら、シーン内容に関して概要文を発話し、簡単な GUI 操作だけで作成するだけで、区間メタデータ、意味メタデータの同時生成が可能となる。

6.4 ライブ番組向けメタデータ生成の作業コスト評価実験

6.4.1 実験概要

前節までに述べた、SceneCabinet / Live! を用いた、ライブ番組向けのメタデータ生成の作業モデルを活用し、ライブ野球中継（全国都市対抗野球）1 試合分の全てのバッターシーンを対象にメタデータを生成する作業のコスト評価実験を行った。具体的には、SceneCabinet / Live! を用いる提案方法と全て手作業で行う方法の作業効率を比較する実験を行った。

実験は複数の作業者によって行い、それぞれの作業モデル毎の作業効率を評価した。以降、におい本節では、SceneCabinet / Live! を用いるメタデータ生成の作業モデルを「作業モデル A」、手作業で全てのメタデータ生成を行う作業を「作業モデル B」と呼ぶこととする。作業モデル A, B において、各バッターシーンに対して、生成するメタデータ項目の具体的な生成ルールを以下に示す。

- 区間メタデータ
  - 開始時間：　ピッチャーがバッターに対して、該当打席の最後のボールを投げ始めるタイミングの時間
  - 終了時間：　該当バッターの打席の結果（例：ヒット、三振等）が判明したタイミングの時間

- 意味メタデータ
  - タイトル：　バッターの名前、打席の結果、イニング数
  - 概要文：　バッターの名前、打席の結果、イニング数の他、アウト数、ランナー出塁状況、ピッチャーの名前を含む文章
  - キーワード：　タイトルと概要文に含まれる名詞と固有名詞

作業モデル A, B 共に 6 名の作業者により実行し、試合中の各バッターのシーンに対するメタデータ生成にかかる時間を計測した。計測に当たっては、番組中で各バッターの打席の結果が判明したタイミングから計測を始め、該当バッターに対する上記の区間メタデータと意味メタデータを生成し終わるタイミングまでの間の時間を計測した。

6.4.2　実験結果

図 6.4.2-1 と図 6.4.2-2 に、6 人の作業者によって、作業モデル A と作業モデル B を実行した結果のグラフをそれぞれ示す。両方のグラフともに、横軸は試合開始からの打者の打席シーンの番号、すなわち、試合における最初の打者、二番目の打者、三番目の打者と続く軸である。縦軸は、該当する打者に対する区間メタデータと意味メタデータの生成にかかった時間を示している。具体的には、ライブ映像にて各打者の結果が判明したタイミングからの遅延時間を示している。図 6.4.2-1 と図 6.4.2-2 共に、6 つのラインが描かれているが、1 つ 1 つのラインが各作業者の作業時間を示すものである。

図 6.4.2-1 は、SceneCabinet / Live! を利用する作業モデル A の結果であり、野球中継が進行しても、6 人全ての作業者について、同一作業者においては、メタデータ生成時間がほぼ一定であったことを示している。1 人の打者に対するメタデータ生成時間として、最長でも作業者 no.4 の 10 分程度の作業時間であっ

た。ライブ映像において、各打者シーンが 10 分以上の長さであったことから、全作業者が試合の進行に遅れることなく、メタデータを生成することができたことを示している。

一方で、手作業でメタデータを生成する作業モデル B では、図 6.4.2-2 に示すように、野球中継の番組が進むにつれ、メタデータ生成にかかる作業時間がどんどん増えていく結果となった。これは、ライブ映像の進行に対し、メタデータの生成作業の時間が追い付かず、ライブ映像からの遅延時間が増えていっている様子が示されたものである。すなわち、ある打者のメタデータを生成する作業を実施している途中で、次の打者のシーンも完了してしまい、作業者は休む間もない状態となったことを示している。具体例として、作業者 no.3 については、13 人目の打者のメタデータ生成は、ライブ映像に対し、約 35 分遅れで完了している。この場合、図 6.2.1-1 でしめした、メタデータ活用したライブ番組の提供にあたって、サービス要件が満たせない可能性が高い状態となる。

以上の結果より、SceneCabinet / Live!を活用する作業モデル A が、人手作業による作業モデル B に比べ、短い作業時間でメタデータを生成できることが確認でき、かつ、メタデータ活用するライブ番組のサービス要件を満たせる可能性が高いことが明らかになった。

作業者一人一人の個人差があるので一概には言えないが、作業モデル A と作業モデル B のメタデータ生成の時間を比較すると、例えば、9 番目の打者シーンに対しては、作業モデル A における 6 人の作業者の平均作業時間は 6 分 17 秒であったのに対し、作業モデル B では 17 分 18 秒であった。これは SceneCabinet / Live! を用いた作業モデルによる作業時間は、手作業のみの作業モデルの時間を 64%削減していることになる。

更には、作業者 no.5 の作業者においては、21 番目の打者シーンのメタデータ生成には、作業モデル A で 3 分 10 秒、作業モデル B で 17 分 1 秒である。作業モデル B の作業時間の約 81%を削減できたことになる。作業者が作業モデル A により慣れてくると、より高いコスト削減効果を示すものと考えられる。

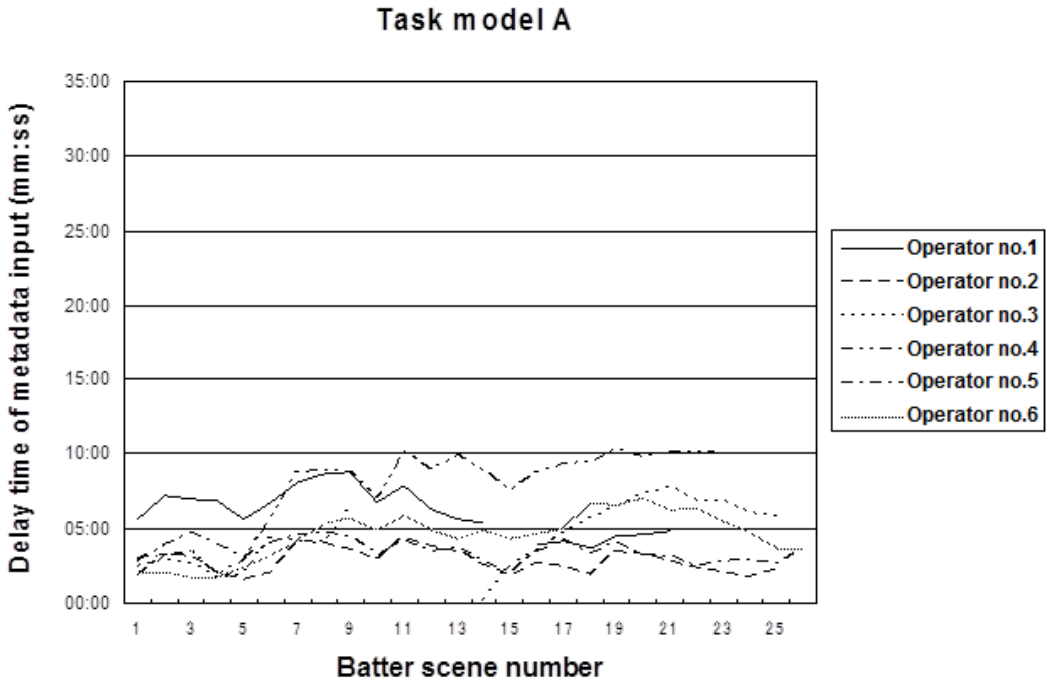


図 6.4.2-1　作業モデル A におけるメタデータ生成時間の遅延時間

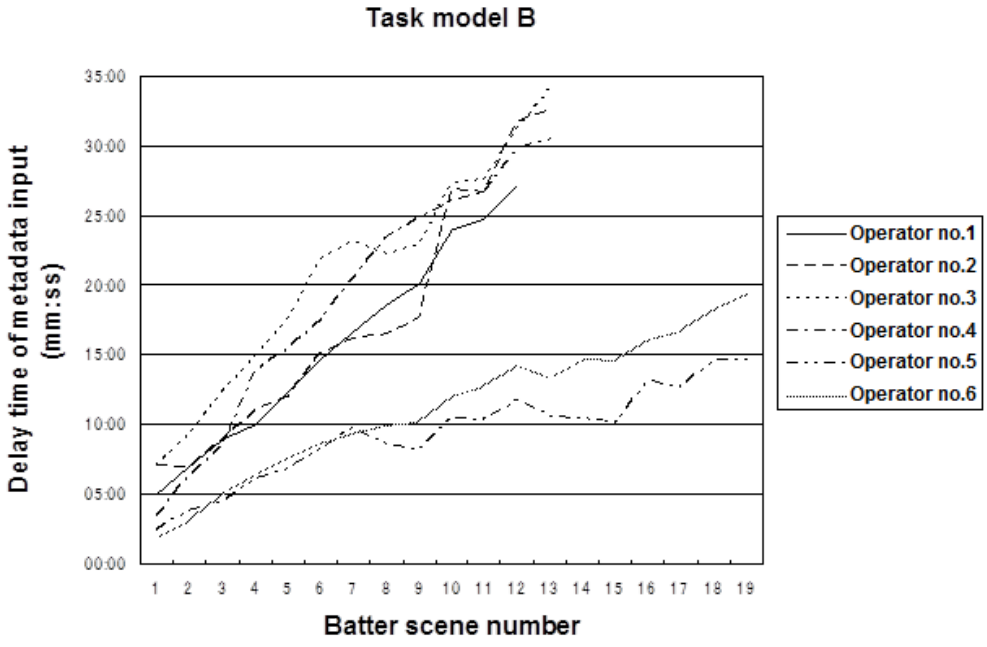


図 6.4.2-2　作業モデル B におけるメタデータ生成の遅延時間

作業者に対しては、2 つの作業モデルに対する疲労感についてのアンケートも行った。作業の疲労感に関して、総じて手作業の作業モデル B のほうが「疲れた」「大変だった」との意見が得られ、SceneCabinet / Live! を利用する作業モデル A は「大変ではなかった」との意見が得られた。特に、概要文の入力作業に関しては、SceneCabinet / Live! を利用した場合は、野球中継の状況から目を離さず、音声を入力するだけで実施できるため、手作業に比べ、非常に効率的にメタデータ生成が実施できたとの意見が得られた。

更に、SceneCabinet / Live! を利用すると、音声認識結果を手掛かりにシーンの頭出しが簡単にできるため、シーンの開始時間、終了時間を探す手間が非常に簡単に行えたとの意見が多く得られた。再生モニタの±1 秒、±10 秒のジャンプボタンも非常に有効であったとの意見も得られた。

6.4.3 考察

以上の結果より、作業者の発話音声の認識を中心とするメディア解析処理、ユーザインタフェースの利用が特徴であるライブ番組向けのメタデータ生成作業モデルのコスト効果について考察する。

本実験では野球中継のバッターシーンに対するメタデータ生成ということで、作業モデル、及び、メタデータ項目のルールも比較的明確なルールを設定することができた。このことは、オペレータが野球に関する専門的な知識を持っていなくても、容易かつ効果的にタスクを実行できたことの理由になると考えられる。

一方、個々のバッターシーンという比較的明確な区切りが存在するシーンではなく、ピッチャー交代、乱闘シーン、監督が指示を出しているシーン等の野球ならではの重要なシーンへのメタデータを生成する場合は、オペレータは、野球のルール等の専門知識も必要となってくる。

しかし、そのような場合であっても、タスクを 2 つの段階に分けることができれば、そのうちの第 1 段階では、ダイジェストで使用されると予想されるすべてのシーンを含む「基本」メタデータの生成と、 最終的なメタデータを作成するための創造的な人材による「基本的な」メタデータと、前のタスクのワークフローを容易に定義することができ、映像認識技術も効果的に適用できる。

したがって、SceneCabinet / Live! によるメタデータ生成は、創造的な人員の負担を軽減し、タスク全体のコスト削減を実現することが十分期待できるものと考えられる。

6.5 まとめ

本章では、ライブ番組向けのメタデータ生成の作業モデルの提案、及び、実験結果と考察を述べた。具体的には、ライブ番組中の見逃したシーン、あるいは、もう 1 度見たいシーンを放送中の好きなタイミングで簡単にダイジェスト視聴するサービスを想定し、その実現に必要なメタデータをリアルタイムで生成する作業モデルの提案、実験結果を述べた。

ライブ野球中継において、各打者のシーンに対するメタデータ生成を行う場合、特にシーンの概要文を「定型文に対する音声認識処理」という高精度の認識処理が実施できる方式を提案した。極力、GUI 操作画面は見ずに、試合の状況だけを見ながら、手間をかけずにメタデータを生成することを意識した方法である。この方式を実装した、ライブ番組向けのメタデータ生成用のユーザインタフェースシステム「SceneCabinet / Live!」を提案した。音声認識結果ブラウザ等の GUI を利用することで、映像の巻き戻し再生などを極力しない形で、メタデータの生成が行えるようデザインしたシステムとなっている。

ライブ野球中継に対し、SceneCabinet / Live! を用いたメタデータ生成の作業コスト削減に関する実験をおこなったところ、6 名の作業者全員が、番組の進行に遅れることなく、メタデータを生成することができることを明らかにした。これを全て人手作業で行うと、番組の進行に対し、メタデータを生成するペースが追い付かなくなり、映像視聴サービスの要件を満たせなくなる可能性が高い。SceneCabinet / Live! を活用した作業モデルは、サービス要件を満たし、市場拡大のチャンスを広げることに寄与できることを明らかにした。

第7章 結論

本研究では、通信ネットワーク上でのシーン単位の映像視聴サービスの市場拡大を狙い、サービスに必要なメタデータ生成の作業コストの低減に関する技術課題に取り組んだ。この技術課題に対し、映像中のテロップ文字の認識技術を始めとする映像・音声認識、自然言語処理といったメディア解析技術により、メタデータの手がかりを自動抽出する方式を提案した。また、自動抽出した結果をメタデータ生成の作業者に分かりやすく提示し、効率的な編集作業を可能とするユーザインタフェース方式の提案を行った。これらの提案方式を実装したシステム「SceneCabinet / NBS」及び「SceneCabinet / Live!」を用いたメタデータ生成の作業モデルを提案し、制作済番組、ライブ番組のメタデータ生成の作業コストとして作業時間がどれだけ短くなるか評価実験を行い、提案方式のコスト削減効果を具体的に示した。

第 2 章では、メタデータ生成のコスト削減にあたっての各種周辺動向と具体的な課題を整理して述べた。映像配信ビジネスを活性化するシーン視聴サービスの実現にあたり、メタデータが重要な情報である点、また、メタデータ生成の作業コストを低減することが産業界から強く求められていたことを述べた。放送事業者での映像制作の業務内容におけるメタデータ生成の位置づけや、人手作業によるメタデータ生成には膨大な時間がかかることを述べた。また、従来技術の動向や到達点、本研究の課題の達成にあったっての問題点を指摘した。最新のファイルベースの映像制作に対し、本研究の課題の意義について明らかにした。

第 3 章では、本研究を進めるにあたっての基本的な考え方とアプローチを述べた。基本的な考え方として、「人手作業の必要性」「作業コストの内容と削減の可能性」「メタデータ生成の効果的な自動化」「ユーザインタフェースのデザイン」の観点を設けて整理して述べた。基本的な考え方に基づき、本研究で提案するメタデータ生成の作業コストを削減する方式は、メディア解析によるメタデータの手がかり情報を自動抽出した後に、その結果を最小限の人手作業で実現するユーザインタフェースを活用してメタデータを完成させる、2 ステップで実現することを述べた。

第 4 章では、メタデータ生成の人手作業を最小限に抑えるための事前処理としてのテロップ文字認識の方式を提案し、実験結果と考察を述べた。従来技術では検討されていない、テロップ文字の視覚的な特徴として、文字の大きさ、表示位置と映像の意味内容との相関に着目した、区間メタデータの自動生成方式を提案した。ニュース番組の各ニューストピックや野球中継の得点シーン等の区間メタデータが従来にない高い精度で自動生成できることを明示した。更に、放送番組全チャンネル分に対し、リアルタイムでテロップ認識処理を行い、ニュース番組のトピック一覧や番組シーンに対するキーワード検索を提供するシステム Telop on Demand を従来にない映像検索システムとして開発したことを述べた。

第 5 章では、制作済の番組映像を対象としたメタデータ生成の作業コストの削減を実現する作業モデルを提案し、実験結果と考察を述べた。作業モデルを実現するメタデータ生成の作業者向けのユーザインタフェースシステムとして「SceneCabinet / NBS」を提案、実装した。SceneCabinet / NBS には、特に、従来、人手作業に時間がかかっていた区間メタデータの生成作業を短縮化する各種機能を備えたものであり、第 4 章で述べたテロップ文字認識方式を含む、複数のメディア解析技術の組み合わせ機能や、キー画像ブラウザ、再生モニタといったメタデータオーサリング GUI がポイントである。実験では、提案作業モデルの作業コスト効果として、ニュース番組全体からニューストピックに関するメタデータ生成について、SceneCabinet / NBS を用いる作業モデルは、全て人手作業で行うモデルの作業時間を約 64%短縮できることを明らかにした。ニュース番組とサッカー中継番組で生成されるメタデータの精度に若干の違いが生じたが、メタデータ生成のルールや映像内容に関する背景知識が重要であることを明らかにした。

第 6 章では、ライブ番組に対し、番組進行中にリアルタイムにメタデータを生成する作業モデルを提案した。メタデータ生成の作業者が極力 GUI 画面ではなく、試合の状況だけを見ながら、簡単にメタデータを作れることを意識し、音声入力を中心とした作業モデルである。この作業モデルを実現するユーザインタフェースシステム「SceneCabinet / Live!」を実装した。ライブ野球中継の各打席シーンに対し、SceneCabinet / Live! を用いたメタデータ生成の作業コスト削減に関する実験を行い、6 名の作業者全員が、番組の進行に遅れなく、メタデータ生成できることを確認した。人手作業では番組の進行に間に合わなくなったのに対し、提案した作業モデルは、新たな市場形成のポテンシャルのある映像視聴サービスの提供に耐えうるレベルであることを明らかにした。

以上、特に、第 2 章から第 6 章は、本研究の課題、従来技術の把握、問題抽

出から提案手法の検討、実験、考察という一連の流れであるが、これら各章の内容が組み合わさり、本研究のテーマである「映像・音声認識・自然言語処理によるメタデータ生成の作業コスト削減」が実現できたといえる。

表 7-1 には、第 4 章、第 5 章、第 6 章の各実験にて、メタデータの各項目が、どのような情報を元に生成されるかの全体像を示す。表中では、特に、メディア解析により自動生成処理は下線を引いて示した。第 4 章では、テロップ文字をベースとしたメディア解析により区間メタデータの候補を自動生成した。第 5 章では、その結果を確認・修正することで区間メタデータを完成させ、意味メタデータはテロップ認識、音声認識、自然言語処理を統合的に活用し、完成させる方式を提案した。第 6 章では、ライブ野球番組に対し、リスピーク音声をベースに区間メタデータ、意味メタデータをほぼ同時に生成する方法を提案した。

表 7-1　第 4 章、第 5 章、第 6 章とメタデータの関係

		第4章			第5章		第6章
		ニュース番組の各トピック	サッカー中継のシュートシーン	野球中継の得点シーン	ニュース番組の各トピック	サッカー中継のシュートシーン	野球中継の各打席シーン
区間メタデータ	開始時間	タイトルテロップとカット点を活用	カメラワークを活用	投手動作を活用	第4章で得られる開始時間の候補を確認・修正	第4章で得られる開始時間の候補を確認・修正	投球動作を活用
	終了時間	タイトルテロップとカット点を活用	CGパターンと選手名テロップを活用	得点テロップを活用	第4章で得られる終了時間の候補を確認・修正	第4章で得られる終了時間の候補を確認・修正	リスピーク音声の入力開始の時間を活用
意味メタデータ	タイトル				ニュースタイトルテロップの文字認識結果を確認・修正	選手名テロップの文字認識結果を確認・修正（スコアシートを確認・修正）	イニング数、選手名、結果を手入力
	概要				音声認識結果を確認・修正（読み原稿を確認・修正）		リスピーク音声の認識結果を確認・修正
	キーワード				タイトル、概要から言語処理で重要語を抽出・選択		タイトル、概要から言語処理で重要語を抽出・選択

表 7-1 で示した各種メディア解析技術、及び、SceneCabinet / NBS と SceneCabinet / Live! の専用 GUI での人手作業により、ニュース、サッカー、野球の各番組について、ライブ番組のような厳しい作業要件においても、従来の人手作業に比べ、大幅なコスト削減が実現できた。ニュース、スポーツの各番組ジャンルに対するメタデータ生成の削減効果を明らかにした本研究の成果は、第 1 章で述べた、番組映像コンテンツをシーン単位で視聴するという、視聴者からの高いニーズがある新しい映像視聴サービスが次々に生み出されること、またそれにより、市場が拡大することに大きく貢献したといえる。

また、2018 年現在の最新の映像制作業務の観点からみると、まず、業務内容が 2007 年の本研究当時とは異なり、ファイルベースという、最初からデジタル形式で行われるようになってきている。すなわち、テロップ文字のテキスト情報など、多くの情報が最初から様々な用途に利用可能なデジタル情報として一元管理されるようになった。一方で、過去の膨大な映像アーカイブは、デジタルではなく、アナログテープの状態のままで管理されている。加えて、ファイルベースにおいても、テロップ文字の大きさや画面上の位置情報といった視覚特徴のデジタル化までは行われていないという状況もある。

すなわち、約 10 年前の本研究成果は、当時としては、テロップ文字の視覚特徴を最大限に利用した初めてのメタデータ生成作業モデルとして、新たな市場創造に大きな貢献をしたが、2018 年の最新の映像制作環境においても、特に、テロップの視覚特徴を活用する点は、今後新たに制作される番組、また、過去番組へのメタデータ生成に対し、有効な方式といえる。

また、本研究の提案方式は、NTT アイティ社（現 NTT テクノクロス社）から、2000 年に「テロップ認識システム」、2005 年に「メディアオーケストラ」として製品化され市場に提供されている。本研究の後も、様々なメディア解析の技術、また、映像配信サービスに関する各種周辺機能（映像アップロード、ユーザ管理等）が追加される等、バージョンアップがなされ、最新状況としては、NTT テクノクロス社の「viaPlatz」として提供が継続されている。このように、本研究の成果は、テロップ認識を積極的に活用し、メタデータ生成の作業を短時間に実行する方式として、初めて市場に出たシステム製品になったと同時に、今現在も viaPlatz の形で長く市場に提供され続けており、市場形成、拡大に貢献をしたといえる。

今後将来的には、本研究における方式検討、実験で得られた知見を基に、ニュースやスポーツ以外の番組ジャンルのメタデータ生成への適用を指向することが考えられる。例えば、ワイドショー、クイズ番組等は、テロップ文字は高頻度に使われるが、近年は、出演者のコメントが全てテロップ文字化されてい

ることから、ニュースやスポーツに比べ、シーンの区切りとして利用することは難しくなることが想定される。一方で、テロップ文字が含まれないドラマ、映画、アニメ等のストーリー性のある番組コンテンツへの適用拡大を考えても、シーン分割する際には、映像内容をより意味レベルまで解析可能なメディア解析技術の検討が必要と考えられる。

これに対する最新の研究動向として、本研究当時に比べ、コンピュータの計算処理能力の飛躍的な向上やビッグデータ活用が進んだことから、ディープラーニングに代表される人工知能（AI）の研究が盛んに行われている。一例として、2016 年に、IBM の Watson は、1 本の映画コンテンツから予告編に相応しい、ストーリー性を考慮した重要なシーン区間だけを抽出・結合し、数分の予告編映像を自動生成する方式を提案している[Smith2]。このように、近年は、技術進歩により、クリエイティブな映像編集に迫るレベルの技術が確立されてきており、今後も高度化が進むと考えられる。しかしながら、映画は勿論、放送番組のような社会的信頼度が高いコンテンツを扱う上では、やはり、最終的には、サービス提供する事前に、人手による最終確認が必要である。すなわち、最新の研究アプローチにおいても、本研究の基本的考え方である、メディア解析と人手作業の組み合わせという 2 ステップの枠組みは適用可能なコンセプトであり、その中の 1 つのステップとして、AI 等によるシーン情報抽出の高度化が図られる構図といえる。

また、人手作業の短時間化には、ユーザインタフェースのデザインが非常に重要であるが、適切なデザインは、番組のジャンルやサービス要件、メタデータ生成の作業モデルの内容によって替わりうると考えられる。業務フローへのインパクトも考慮しつつ、最適なデザインを検討することが新たな研究課題として考えられる。今後、これらの課題へのチャレンジを進め、更なる産業貢献を指向していく。

謝辞

本論文の作成にあたり、筑波大学大学院博士後期課程在籍時に、懇切丁寧なご指導を頂きました、筑波大学大学院システム情報工学研究科知能機能システム専攻の亀田能成教授に感謝の意を表します。

本論文の副査をお引き受け頂き、また適切なご意見を頂きました、筑波大学大学院システム情報工学研究科知能機能システム専攻の宇津呂武仁教授、北原格准教授、東京理科大学の谷口行信教授、日本工業大学の新井啓之教授に、感謝の意を表します。また、良い研究環境を提供して頂き、分け隔てなく接していただいた、画像情報研究室のメンバの皆様に感謝致します。

本研究は、筆者が日本電信電話株式会社(NTT)において入社以来行ってきた様々な研究成果を基に執筆致しました。その中で、映像メディア解析に関するご指導を頂いた、元筑波大学図書館情報メディア研究科の小高和己教授(元 NTT ヒューマンインタフェース研究所主幹研究員)、元株式会社 NTT ドコモマルチメディア研究所の倉掛正治部長（元 NTT ヒューマンインタフェース研究所主任研究員)、NTT アドバンステクノロジ株式会社の児島治彦ユニット長（元 NTT サイバーソリューション研究所主席研究員)、NTT テクノクロス株式会社の仲西正マネージャ（元 NTT サイバーソリューション研究所主幹研究員）に感謝致します。

また、本研究の遂行にあたり、放送通信連携サービス、メタデータ技術に関するご指導を頂いた、NTT サービスイノベーション総合研究所の川添雄彦所長、NTT サービスエボリューション研究所の阿久津明人主席研究員、NTT テクノクロス株式会社の山田智一担当部長（元 NTT サービスエボリューション研究所主幹研究員)、NTT アドバンステクノロジ株式会社の松尾義博担当部長（元 NTT メディアインテリジェンス研究所主幹研究員）に感謝致します。

また、日頃の会社業務でのご支援に加え、社会人としての博士号取得に向けてご理解と励ましを頂いた、NTT 西日本研究開発センタの髙橋郁也所長、及びチームメンバに感謝の意を表します。

最後に、本論文の執筆を支えてくれた、私の家族に感謝致します。

参考文献

[TV-Anytime] TV-Anytime Forum ウェブサイト, http://www.tv-anytie.org/

[ETSI] ETSI Technical Specification 102 822 - 3 - 1 v1.10.1, "TV-Anytime Metadata schemas".

[ARIB] 電波産業会ARIB技術資料 TR-B38, "VHF-Low帯に適用するセグメント連結伝送方式による地上マルチメディア放送運用規定".

[Mackay] W.E.Mackay, "EVA: an experimental video annotator for symbolic analysis of video data," ACM SIGCHI Bulletin, Vol.21, Issue2, pp.68-71, 1989.

[Mackay2] W.E.Mackay and G.Davenport, "Virtual Video Editing in Interactive Multimedia Applications," Communications of the ACM, Vol.32, No.7, pp.802-810, 1989.

[Ueda] 上田博唯, 宮武孝文, 吉沢聡, "認識技術を応用した対話型映像編集方式の提案," 信学論(D-II), Vol.J75-D-2, No.2, pp.216-225, 1992.

[Nagasaka] 長坂晃朗, 田中譲, "カラービデオ映像における自動索引付け法と物体検索法," 情処学誌, Vol.33, No.4, pp.543-550, 1992.

[Hjelsvold] R.Hjelsvold, "Integrated video archive tools," Proc. of ACM Multimedia '95, pp.283-293, 1995.

[Kanade] 金出武雄, 佐藤真一, "Informedia: CMU ディジタルビデオライブラリプロジェクト," 情報処理, Vol.37, No.9, pp.841-847, 1996.

[Satoh] S.Satoh, Y.Nakamura, and T.Kanade, "Name-It: Naming and detecting faces in news videos," IEEE Multimedia, vol.6, no.1, pp.22-35, 1999.

[Nitta] 新田直子, 馬場口登, 北橋忠宏, “放送型スポーツ映像の構造を考慮した重要シーンへの自動アノテーション付け,” 信学論(D-II), Vol.J84-D-2, No.8, pp.1838-1847, 2001.

[Taniguchi] 谷口行信, 南憲一, 佐藤隆, 桑野秀豪, 児島治彦, “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(D-II), Vol.J84-D2, No.6, pp.1112-1121, 2001.

[Sumiyoshi] 住吉英樹, 佐野雅規, 八木伸行, “メタデータ制作フレームワーク,” 映情学誌, Vol.61, No.2, pp.152-157, 2007.

[MPEG] MPEG (The Moving Picture Experts Group) の MPEG7 標準仕様に関するウェブサイト, https://mpeg.chiariglione.org/standards/mpeg-7

[Mediasite] Mediasite Enterprise Video Platform, メディアサイト株式会社, http://www.mediasite.co.jp/products/platform.

[Horibuchi] 堀淵惣一郎, “ファイルベースワークフローの現状と今後,” 映情学誌, Vol.71, No.4, pp.444-450, 2017.

[Smith] M.A. Smith and T.Kanade, “Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques,” Proc. of CVPR’97, pp.775-781, 1997.

[Gargi] U.Gargi, “A System for Automatic Text Detection in Video,” Proc. of 5th ICDAR, pp.29-32, 1999.

[Ariki] Y.Ariki and K.Matsuura, “Automatic Classification of TV News Articles based on Telop Character Recognition,” Proc. of IEEE Multimedia Systems’99, pp.148-152, 1999.

[Lienhart] R.Lienhart, “Automatic Text Recognition for Video Indexing,” Proc. of ACM Multimedia’96, pp.11-20, 1996.

[Shim] J.Shim, “Automatic Text Extraction from Video for Content-Based Annotation and Retrieval,” Proc. of ICPR’98, pp.618-620, 1998.

[Hori] 堀修, 三田雄志, “テロップ認識のための映像からのロバストな文字部抽出法,” 信学論(D-II), Vol.J84-D2, No.8, pp.1800-1808, 2001.

[Mori] 森稔, 倉掛正治, 杉村利明, 塩昭夫, 鈴木章, “背景・文字の形状特徴と動的修正識別関数を用いた映像中テロップ文字認識,” 信学論(D-II), Vol.J83-D2, No.7, pp.1658-1666, 2000.

[Kawai] 河合吉彦, 住吉英樹, 八木伸行, “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,” 信学論(D), Vol.91-D, No.10, pp.2529-2539, 2008.

[Kumano] 熊野雅仁, 有木康雄, 上原邦昭, 下條真司, 春藤憲司, 塚田清志, “映像編集支援システムのためのショットサイズ自動付与,” 信学論(D-I), Vol.J85-D1, No.7, pp.592-602, 2002.

[Suzuki] 鈴木賢一郎, 中嶋正臣, 坂野鋭, 三部靖夫, 大塚作一, “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-II), Vol.J86-D2, No.4, pp.468-478, 2003.

[Miura] 三浦宏一, 高野求, 浜田玲子, 井手一郎, 坂井修一, 田中英彦, “料理映像の構造解析による調理手順との対応付け,” 信学論(D-II), Vol.86-D2, No.11, pp.1647-1656, 2003.

[Takimoto] 瀧本政雄, 佐藤真一, 坂内正夫, “大容量放送映像アーカイブからの同一フラッシュシーン映像の発見,” 信学論(D), Vol.J89-D, No.12, pp.2699-2709, 2006.

[Andou] 安藤彰男, 今井亨, 小林彰夫, 本間真一, 後藤淳, 青山信正, 三島剛, 小早川健, 佐藤庄衛, 尾上和穂, 世木寛之, 今井篤, 松井淳, 中村章, 田中英輝, 都木徹, 宮坂栄一, 磯野春雄, “音声認識を利用した放送用ニュース字幕制作システム,” 信学論(D-II), Vol.J84-D2, No.6, pp.877-887, 2001.

[Bessho] 別所克人, 松永昭一, 大附克年, 廣嶋伸章, 奥雅博, 林良彦, “話題構造抽出に基づく会議音声インデクシングシステム,” 信学論(D), Vol.J91-D, No.9, pp.2256-2267, 2008.

[Kobayashi] 小林彰夫, 奥貴裕, 本間真一, 佐藤庄衛, 今井亨, 都木徹, “単語誤り最小化に基づく識別的リスコアリングによるニュース音声認識,” 信学論(D), Vol.J93-D, No.5, pp.598-609, 2010.

[Fujimoto] 藤本雅清, 有木康雄, 松本宏, “音声情報と画像情報の併用による商品紹介映像のセグメンテーション,” 信学論(D), Vol.J89-D, No.2, pp.292-304, 2006.

[Yoshida] 吉田壮, 小川貴弘, 長谷川美紀, “歌謡番組における映像の構造に着目したシーン分割手法,” 信学論(D), Vol.J97-D, No.7, pp.1177-1188, 2014.

[Mikami] 三上弾, 紺谷精一, 森本正志, “突発音検出と教師なし動き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,” 信学論(D), Vol.J90-D, No.2, pp.526-534, 2007.

[Nemoto] 根本敦史, 半谷精一郎, 宮内一洋, “テロップの認識による資料映像の検索について,” 1994 春季信学全大, D-427.

[Nakajima] 中島康之, 堀裕修, 塩原敏充, “キーワード画像抽出による動画像サマリの作成,” 1994 情処 49 回全大, 2-91.

[Satoh] 佐藤隆, 新倉康巨, 谷口行信, 阿久津明人, 外村佳伸, 浜田洋, “MPEG 符号化映像からの高速テロップ領域検出法,” 信学論 (D-II), vol.J81-D2, no.8, pp.1847-1855, 1998.

[Takano] 高野正次, 中村修, “H.261 符号ハンドリングによるテロップ検出,” 画像電子学会研究会予稿, 94-06-04, pp.13-16, 1995.

[Mogi] 茂木祐治, 有木康雄, “ニュース映像中の文字認識に基づく記事の索引づけ,” 信学技報, PRU95-240, 1996.

[Kurakake] S. Kurakake, H. Kuwano, and K. Odaka, “Recognition and visual feature matching of text region in video for conceptual indexing,” IS&T/SPIE Electrical Imaging ’97, vol.3022, pp.368-379, 1997.

[Zhong] Yu Zhong, Kalle Karu, and Anli K.Jain, “Locating Text in Complex Color Images,” IEEE ICDAR, pp.146-149, 1995.

[Minami] 南憲一, 阿久津明人, 浜田洋, 外村佳伸, “音情報を用いた映像インデクシングとその応用,” 信学論(D-II), Vol.J81-D2, No.3, pp.529-537, 1998.

[Ohtsuki] K.Ohtsuki, T.Matsuoka, S.Matsunaga, and S.Furui, “Topic Extraction based on Continuous Speec Recognition in Broadcast-news Speech,” Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, pp.527-534, 1997.

[Hayashi] Y.Hayashi, K.Ohtsuki, K.Bessho, O.Mizuno, Y.Matsuo, S.Matsunaga, M.Hayashi, T.Hasegawa, and N.Ikeda, “Speech-based and Video-supported Indexing of Multimedia Broadcast News,” Proc. of SIGIR, pp.441-442, 2003.

[Smith2] J.R.Smith, D.Joshi, B.Huet, W.H.Hsu, and Jozef Cota, “Harnessing A.I. for Augmenting Creativity: Application to Movie Trailer Creation,” Proc. of ACM Multimedia 2017, 2017.

公表論文リスト

【査読付き学術論文誌】

[J1] 桑野秀豪, 松尾義博, 川添雄彦, “映像・音声認識, 自然言語処理の適用によるメタデータ生成の作業コスト削減効果に関する考察,” 映像情報メディア学会論文誌, Vol.61, No.6. pp.842-852, 2007.

[J2] H.Kuwano, Y.Kon’ya, T.Yamada and K.Kawazoe, “SceneCabinet Live!: Generation of Semantic Metadata Combining Media Analysis and User Interface Technologies,” SMPTE Motion Imaging Journal, Vol. 114, Issue.12, pp.446-452, 2005.

【査読付き国際会議論文】

[C1] H.Kuwano, S.Kurakake and K.Odaka, “Telop Character Extraction from Video Data,” Proceedings of Document Image Analysis 97, pp.82-88, 1997.

[C2] H.Kuwano, Y.Taniguchi, H Arai, M.Mori, S.Kurakake, and H.Kojima, “Telop on Demand: Video Structuring and Retrieval based on Text Recognition,” Proceedings of IEEE International Conference on Multimedia and Expo (ICME) 2000, pp.759-762, 2000.

[C3] H.Kuwano, Y.Taniguchi, K.Minami, M.Morimoto and H.Kojima, “A smart TV viewing and Web access interface based on video indexing techniques,” Proceedings of IEEE International Conference on Consumer Electronics (ICCE) 2002, pp.204-205, 2002.

[C4] H.Kuwano, Y.Matsuo and K.Kawazoe, “SceneCabinet : Semantic Metadata Extraction System combining Video/Audio Indexing and Natural Language Processing Techniques,” Proceedings of International Broadcasting Convention (IBC) 2004, pp.458‐466, 2004.

[C5] 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,” Proceedings of International Broadcasting Convention (IBC) 2005, pp.253‐260, 2005.

【国内外発表】

[P1] 谷口行信, 南憲一, 佐藤隆, 桑野秀豪, 児島治彦, “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(D‐II), Vol.J84‐D2, No.6, pp.1112‐1121, 2001.

[P2] 新井啓之, 桑野秀豪, 倉掛正治, 杉村利明, “映像中のテロップ表示フレーム検出方法,” 信学論(D-II), Vol.J83‐D2, No.6, pp.1477-1486, 2000.

[P3] Y.Kon’ya, H.Kuwano, T.Yamada, M.Kawamori and K.Kawazoe, “Metadata Generation and Distribution for Live Programs on Broadcasting-Telecommunication Linkage Services,” Proceedings of Advances in Multimedia Information Processing‐PCM 2005, Vol.3767, pp.224‐233, 2005.

[P4] S.Kurakake, H.Kuwano and K.Odaka, “Recognition and Visual Feature Matching of Text Region in Video for Conceptual Indexing,” Proceedings of SPIE3022, Storage and Retrieval for Image and Video Databases V, 368, 1997.

[P5] 桑野秀豪, 山田智一, 川添雄彦, “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,” 信学技報, vol.105, no.674, PRMU2005‐302, pp.283‐288, 2006.

[P6] 桑野秀豪, 倉掛正治, 小高和己, “映像データ検索のためのテロップ文字抽出法,” 信学技報, PRMU1996-385, pp.39-46, 1996.

[P7] 新井啓之, 桑野秀豪, 倉掛正治, 小高和己, “映像中被写体検索のための部品抽出方法の検討,” 信学技報, PRMU1996-563, pp.23-29, 1997.

[P8] 桑野秀豪, 新井啓之, 倉掛正治, 杉村利明, “映像中に挿入された部分画像検出方法,” 信学全大, 情報システム(2), 183, 1999.

[P9] 新井啓之, 桑野秀豪, 倉掛正治, 小倉健司, “映像中の流れるテロップ文字列の抽出方法,” 信学全大, 情報システム(2), 234, 1998.

[P10] 新井啓之, 桑野秀豪, 倉掛正治, 杉村利明, “映像中の静止/ロールテロップの検出法,” 信学ソ大, 267, 1998.

[P11] 倉掛正治, 新井啓之, 桑野秀豪, 杉村利明, “PC版映像中テロップ認識システム,” 信学ソ大, 265, 1998.

[P12] 新井啓之, 桑野秀豪, 倉掛正治, 小高和己, “映像中被写体検索のための部品抽出方法,” 信学全大, 情報システム(2), 266, 1997.

[P13] 桑野秀豪, 倉掛正治, 小高和己, “カラー画像からの高速テロップ文字領域抽出法,” 信学全大, 情報システム(2), 265, 1997.

[P14] 倉掛正治, 桑野秀豪, 小高和己, “テロップ情報自動インデクシングシステムリアルタイム版,” 信学全大, 情報システム(2), 264, 1997.

[P15] 新井啓之, 桑野秀豪, 倉掛正治, 小倉健司, “色の変化にロバストな被写体抽出方法の検討,” 信学ソ大, 216, 1997.

[P16] 倉掛正治, 桑野秀豪, 新井啓之, 小高和己, “認識技術を用いた映像中キーターゲットインデクシングの検討,” 信学技報, IE1995-582, pp.15-20, 1996.

[P17] 倉掛正治, 桑野秀豪, 新井啓之, 安部伸治, 小高和己, “映像検索のためのテロップ情報自動インデクシング,” 信学ソ大, 547, 1996.

[P18] 石井晋司, 伊藤宏一, 桑野秀豪, “携帯端末向けマルチメディア放送におけるアクセス制御技術,” NTT技術ジャーナル, 23-5, pp.24-27, 2011.

[P19] 石井晋司, 内田良隆, 森住俊美, 松井龍也, 伊藤宏一, 桑野秀豪, 阿久津明人, 関野公彦, “ISDB-Tmmにおけるコンテンツ保護とアクセス制御技術,” 情処技報, AVM-71-11, pp.1-5, 2010.

[P20] 前橋佳林, 桑野秀豪, 谷口行信, 阿久津明人, “ライフスタイルと地理的特徴との関係を利用した場所メタデータ自動生成,” 情処全大, ネットワーク, pp.19-20, 2010.

[P21] 前橋佳林, 桑野秀豪, 谷口行信, 阿久津明人, “来訪者特徴の推定を利用した「場所メタデータ」自動生成,” 信学技報, 109-450, pp.49-54, 2010.

[P22] 桑野秀豪, 紺家裕子, 山田智一, 川添雄彦, “ライブ番組向けダイジェスト視聴サービスのためのリアルタイムメタデータ生成技術,” NTT技術ジャーナル, 17-6, pp.14-17, 2005.

[P23] 紺家裕子, 桑野秀豪, 山田智一, 川添雄彦, “メタデータ生成システムのライブ放送への適用について,” 信学全大, 情報システム(2), D-11-73, 2005.

[P24] 桑野秀豪, 松尾義博, 川添雄彦, “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,” NTT技術ジャーナル, 16-5, pp.22-25, 2004.

[P25] 桑野秀豪, 谷口行信, 児島治彦, “二段階ライン二値化による低解像度テロップ文字領域抽出,” 信学ソ大, D-12-20, 207, 2000.

[P26] 桑野秀豪, 谷口行信, 児島治彦, “映像構造化のためのテロップ属性抽出・分類法の提案,” 情処全大, pp.313-314, 1999.

[P27] 桑野秀豪, “次世代Web技術HTML5の動向とコンテンツ流通サービスへのインパクト,” 信学技報, IN112-307, pp.25-26, 2012.

[P28] 桑野秀豪, 新井啓之, 倉掛正治, 小倉健司, “色劣化に対処する映像中文字領域の抽出方法,” 信学全大, D12-34, 1998.

[P29] 桑野秀豪, 新井啓之, 倉掛正治, 杉村利明, “ライン単位の二値化による映像中文字領域の抽出方法,” 信学ソ大, D-12-42, 1998.

[P30] 紺家裕子, 秋野泰志, 堀口恭太郎, 川森雅仁, 桑野秀豪, 山田智一, 川添雄彦, “IPTVにおけるメタデータ活用技術,” NTT技術ジャーナル, 2006年8月号, pp.19-22, 2006.