

## アノテーションに基づくビデオ検索システムの提案

大平 茂輝†      長尾 確‡      白井 克彦†

† 早稲田大学理工学部

‡ 名古屋大学工学部

〒 169-8555 東京都新宿区大久保 3-4-1 55S-609

E-mail: ohira@shirai.info.waseda.ac.jp

あらまし 映像や音声を含むマルチメディアコンテンツは、テキストコンテンツに比べて、内容に基づく処理が極めて困難である。そこで、マルチメディアコンテンツの検索・変換を行う上で必要となるインデックス情報を生成・加工し、これらアノテーションと呼ばれるメタ情報に基づいたマルチメディアコンテンツの高度利用について、ビデオデータを対象に研究を進めている。本研究では、我々が目指しているマルチメディアコンテンツ検索システムの概要について説明し、アノテーションに基づいたビデオ検索を実現する手法と解決すべき問題について、現状のシステムを例に提案する。

キーワード マルチメディアコンテンツ, ビデオ検索, アノテーション

## A Proposal of Annotation-Based Video Retrieval System

Shigeki Ohira†      Katashi Nagao‡      Katsuhiko Shirai†

†School of Science and Engineering, Waseda University

‡School of Engineering, Nagoya University

3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan

E-mail: ohira@shirai.info.waseda.ac.jp

**Abstract** This paper proposes a tool for multimedia annotation and a video retrieval system based on the annotation. The annotation tool allows users to easily create annotation data including video transcripts using speech recognition, video scene descriptions, and visual object descriptions. The annotation data is used in the multimedia content retrieval and preference-based content transformation. The annotation data is described using XML. This paper also cites and considers some problems in our annotation-based video retrieval system such as amount, accuracy and time cost of content description.

**Key words** Multimedia Contents, Video Retrieval, Annotation

# 1 はじめに

映像や音声を含むマルチメディアコンテンツは、テキストコンテンツに比べて、内容に基づく処理が極めて困難である。そこで、マルチメディアコンテンツの検索・変換を行う上で必要となるインデックス情報を生成・加工し、これらアノテーションと呼ばれるメタ情報に基づいたマルチメディアコンテンツの高度利用について、ビデオデータを対象に研究を進めている [1][2][3]。

ビデオデータが異なる環境で利用される状況を考える場合、その対処方法としては以下の2通りが考えられる。

- 異なる環境ごとに合わせたデータをあらかじめ用意しておく
- オリジナルデータを環境に合わせて変換する

現在のインターネットにおける Web ページの閲覧については、例えば PC 用、モバイル (PDA 等) 用、携帯電話用のページを別々に用意するというように、前者の方法がとられていることが多い。我々が目標とするのは後者の方法である。すなわち、コンテンツ提供者はデータを 1 種類のみ用意すればよく、利用者側の環境や要求に応じてサーバがコンテンツ変換を行うようなシステムである。アノテーションに基づくさまざまなコンテンツ加工を総称して、セマンティック・トランスコーディングと呼ぶ [4]。

ただし、これを可能にするためには、コンテンツ変換を容易にするアノテーションを適宜付与しておく必要がある。コンテンツ変換を容易にするアノテーションとは、すなわちそのコンテンツの内容記述である。ビデオデータであれば、データ中のシーンやシーンに含まれる物体、人物の発話内容などである。

本稿では、我々が目指しているマルチメディアコンテンツ検索システムの概要について説明し、アノテーションに基づいたビデオ検索を実現する手法を現状のシステムを例に提案、解決すべき問題について考察する。

## 2 先行研究・事例

完全なビデオデータを対象とした研究や事例は少ないが、単語と動画の相互検索 [5]、話者と発話

内容の同時検索 [6]、Web 上の文書と画像のクロスメディア検索 [7]、画像から画像の検索、ニュース音声のトランスクリプトに対する検索 [8] など、これまでに多くの研究がなされており、岡らのグループが研究・開発した CrossMediator [9] は、現在実用化されているマルチメディア検索システムの 1 つである。彼らは、音素や濃度ヒストグラムといったデータの時系列特徴量をインデックスとして用いて検索を行っている。

音声データを音素系列として記述する方法は、音声認識誤りによる検索精度の劣化や未知語に対して強いという利点を持つ一方で、同音語や単語境界誤り、短い単語による精度劣化や、内容に基づいた検索や要約が難しいという欠点がある。

個人が家庭で録画したデータに対して検索を行う場合には、このような手法が適していると考えられるが、Web 上で無数のデータが公開されるような場合、それらに対して検索・要約を行うためには十分な内容記述が必要であると考えられる。

## 3 ビデオ検索におけるアノテーションの有用性

### 3.1 テキスト検索・イメージ検索との比較

例として、次のような検索要求を考えてみる「テロで航空機がビルに激突したらいい」

テキスト検索ではどうだろうか。Google [10] で And をとる単語を増やしながら検索した結果を表 1 に示す。

表 1: Google によるテキスト検索例

検索キーワード	ヒット件数
テロ	376,000
テロ, ビル	76,500
航空機, ビル	30,700
テロ, 航空機	25,500
テロ, 航空機, ビル	11,400
貿易センタービルに航空機が激突	90

検索キーワードとしてユーザが与える語数は平均 1, 2 語と言われているが、テキストコンテンツの量がビデオコンテンツとは比較にならないほど多いことを考慮しても、検索結果としてユーザが確認・視聴するためには、相当の絞り込みや提示方法の工夫が必要であることが予想される。

一方、イメージ検索について考えてみると、この例の場合「航空機」「ビル」「爆発・炎上」といったイメージあるいは単語列から検索を行うことになる。

イメージからの検索は精度面から考えて現状では非常に困難であるため、ここでは単語列からの検索のみについて考える。同様に Google のイメージ検索を利用した結果を表 2 に示す。

表 2: Google によるイメージ検索例

検索キーワード	ヒット件数 (正解)
テロ	1,700 (計測不能)
テロ, ビル	117 (15)
航空機, ビル	44 (6)
テロ, 航空機	37 (6)
テロ, 航空機, ビル	11 (4)

上記の正解件数は、テロに関係するものをカウントしており、この中で実際に航空機が写っているものはわずかであった。これは、イメージの説明文としてニュース記事の文章を利用していることにより、航空機という単語が補完された結果であると考えられる。逆に、ニュースのように公共性の高い情報でなく十分な説明がなされていない場合には、イメージの検索は容易ではない。これはビデオ検索についても同様に当てはまることである。

### 3.2 ビデオデータの特徴

ビデオデータの最大の特徴は、映像が持つ過去の事象の偽らざる連続性である。テキストには、それを扱う人間の知識や状況が反映され、時間軸の前後やスキップも容易である。イメージには、瞬間の情報は凝縮されているが、時間情報が欠落している。

ビデオデータをテキスト (実際には音声) とイメージの合成と考えると、補完の最も難しい情報は時間の経過によって我々が得ることのできる真実である。

先ほどのテロの例を考えてみる。同様の例として、イタリア・ミラノで起きた小型機事故と合わせて、テロという事実情報に着目してみる。

米国同時多発テロの場合は、事件の初期段階では、テロという言葉は断定的には使われていない。大統領の声明により、その瞬間から事件はテロであるとの認識が定まったわけである。このため、1 機目と 2 機目の激突のシーンに対してテロという情報を与えるには、その真実を理解した人間の介在が必要になる。

一方、イタリア・ミラノ小型機事故の場合は、事件の第一報で「テロ攻撃の可能性が非常に高い」との見方が示されたと報じたが、その後に事故説と自殺説が出た。このため、事件直後のテロという情報が誤っていることを与えるためには、やはり同様に人間の介在が必要である。

このように、ビデオデータの場合は、必ずしもその時点での音声や映像が真実を表しているとは限らない。そのため、後からアノテーション情報を付与することが可能な仕組みが非常に重要であると考えられる。

### 3.3 アノテーションを利用したコンテンツ変換例

アノテーションが付与されているデータは、様々な自然言語処理が可能である。その例として、ビデオデータから HTML ドキュメントへのコンテンツ変換を行った例を図 1, 2 に示す。

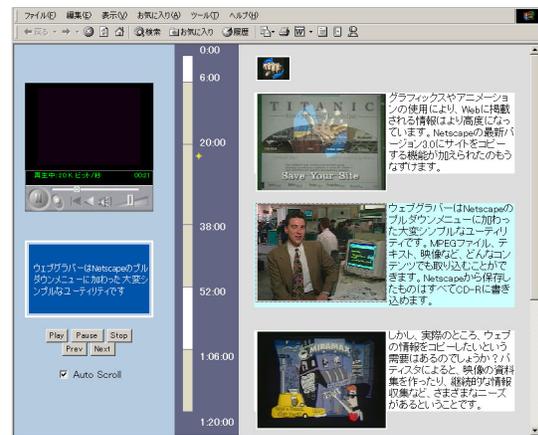


図 1: ビデオドキュメント



図 2: 携帯端末用に変換されたビデオドキュメント

このように、検索されたビデオデータを、ユーザの利用環境や嗜好に応じて適宜変換して提示することも、アノテーションを付与することによって可能になる。一度に多くの検索結果を視聴することのできないビデオデータにとって、これらの処理を可能にするアノテーション情報は必要不可欠である。

## 4 アノテーションに基づくビデオ検索システム

### 4.1 システム概要

我々が提案するアノテーションに基づいたビデオ検索システムの概要を図3に示す。

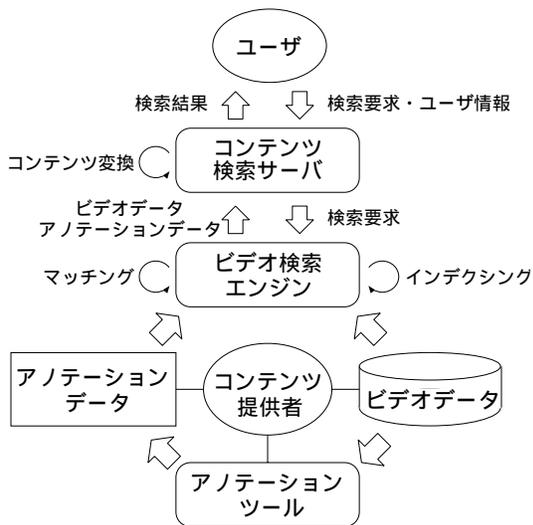


図3: ビデオ検索システムの概要

システムは、コンテンツ提供者側によるアノテーション生成処理、ユーザからの検索要求に対する検索処理、ユーザ環境に応じた検索結果の変換処理、の大きく3つに分けられる。

検索結果の変換処理については、セマンティック・トランスコーディング [4] によって行う。本技術に関する説明は他の文献に譲り、本稿では詳述しない。

### 4.2 アノテーション生成処理

我々は、ビデオデータ中の発話情報、シーン情報、シーン内オブジェクト情報をアノテーションとして生成・編集・関連付けすることを可能にするツールとして、多言語ビデオトランスクリプターを開発している (図4)。



図4: 多言語ビデオトランスクリプター

ビデオデータへのアノテーションの付与は、図5に示される手順にしたがって行われ、作成したアノテーションデータは、XML [11] で記述される。

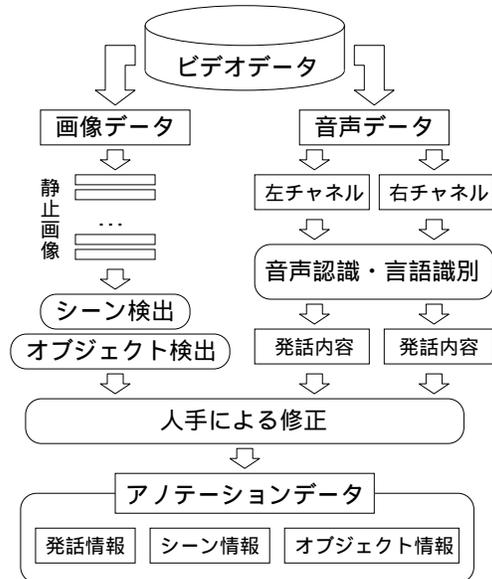


図5: ビデオアノテーションデータ生成までの手順

アノテーションによって扱う内容記述は、シナリオにおけるト書きのような情景描写 (シーン) と登場人物の台詞に相当する発話文 (トランスクリプト)、およびフレーム内に登場するオブジェクトの記述から構成される。

発話情報にはタイムコードと発話内容、シーン情報にはタイムコードとインデックスタイトル、オブジェクト情報にはタイムコードと名称、説明、矩形、リンク情報等が含まれる。

現在、音声認識、言語識別、シーン検出は自動的に処理されているが、認識誤り訂正、シーン統合、オブジェクトの切り出し、シーン記述等は、人間が行う仕組みになっている。

### 4.3 アノテーションに基づく検索処理

生成されるアノテーションデータには、代表シーン画像やオブジェクト画像も含まれるが、これらは検索後のコンテンツ変換や要約時に使われ、検索時には利用されない。検索の原理としては通常のテキスト検索と同じであり、代表的なベクトル空間モデルを用いて類似度を計算する。

ビデオデータの場合は、類似度の高いデータの中から目的のシーン(タイムコード)を検出することが要求されるため、類似度計算はビデオファイル全体とビデオファイル中に含まれる全シーンに対して行われる。

ビデオファイル  $V_i$  が  $m$  個のシーンから構成されるとする。

$$V_i = \{S_1, S_2, \dots, S_m\}$$

このとき、各シーンの特徴ベクトルを  $\overrightarrow{V_i(S_j)}$  とすると、ビデオ  $V_i$  の特徴ベクトルは、シーンの時間長とターム数によって正規化された次の式で表される。

$$\overrightarrow{V_i} = \sum_{j=1}^m \frac{l_j}{L} \cdot \frac{n_j}{N} \cdot \overrightarrow{V_i(S_j)}$$

( $l_j$ : シーン時間長,  $L$ : ビデオ時間長,  
 $n_j$ : シーン中ターム数,  $N$ : 全ターム数)

各シーンの特徴ベクトル  $\overrightarrow{V_i(S_j)}$  は、発話情報  $S_j(t)$ , シーン情報  $S_j(s)$ , オブジェクト情報  $S_j(o)$  中に出現するタームのタームベクトルと、各々の情報に対する重み係数  $\alpha, \beta, \gamma$  によって以下の式で表される。

$$\overrightarrow{V_i(S_j)} = \alpha \cdot \overrightarrow{S_j(t)} + \beta \cdot \overrightarrow{S_j(s)} + \gamma \cdot \overrightarrow{S_j(o)}$$

$$(0 \leq \alpha, \beta, \gamma \leq 1)$$

タームベクトルを構成する各タームの重みは、TF-IDF 法によって求められる。

検索は、まず全ビデオファイルに対してビデオの特徴ベクトルを用いて類似度計算を行い、次にその中の上位  $R$  件に含まれる全シーンに対して同様に類似度計算を行いランク付けする。最終的に、ランキング上位の  $r$  シーンを検索結果として出力する。

## 5 システムの問題点

多言語ビデオトランスクリプターにより、人間の介入する半自動的なビデオアノテーション作成の支援が可能になったが、アノテーションを付与する際の問題として、次の2点が挙げられる。

- 内容記述量、精度
- 内容記述コスト

これらは、処理の自動化と検索・要約等の実用性の意味において互いにトレードオフの関係にある。そこで、以下ではそれぞれの問題を解決する上で今後取り組む必要のある技術的課題について考察する。

### 5.1 内容記述量・精度

すでに述べたとおり、マルチメディアコンテンツは内容に基づく処理が極めて困難であることから、アノテーションツールを用いた内容記述は、ビデオ等のコンテンツの検索・要約を実現するために重要な役割を果たすと考える。しかし、内容記述が乏しかったり、機械処理に頼ることで内容記述の精度が悪いと、その後の検索や要約結果の精度を低下させる原因となる。

図6に示すグラフは、音声認識誤りをシミュレーションして、実際の検索精度にどの程度影響を及ぼすかを分析するとともに、検索において期待されるアノテーション情報の質を予測したものである。

具体的には、RWCP 検索・要約用ニュース音声データベース(192記事, 3,517異なり単語)[12]を用いて、検索対象となる記事に含まれる単語をランダムに除去した場合と、文書の特徴付けに有効でない単語をIDF(逆文書頻度)値により優先的に除去した場合の2通りについて検索実験を行い比較した。

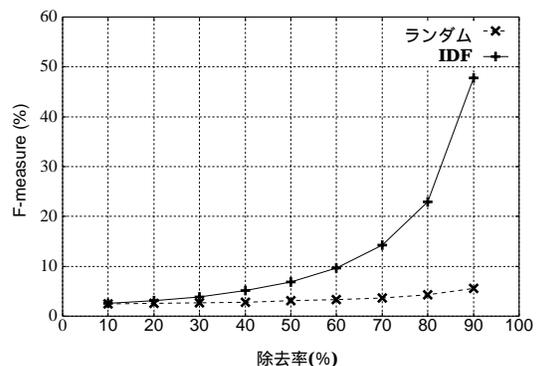


図6: 単語除去率と検索精度の関係

評価尺度には、再現率と適合率から求められる F-measure を用いた。入力クエリーにオリジナルの記事を用いており検索対象数も少ないことから、除去率が小さいときには両者に大きな差は見られないが、除去率 90% になるとその差は 40% 程度になる。すなわち、検索に有意な 10% 以上の単語を音声認識の脱落誤りによって失うと、検索精度に多大な影響があることがわかる。さらに、置換誤りや挿入誤りが起こると検索精度に対する信頼性は著しく低下する。

そこで、内容記述量を向上させると同時に内容記述精度そのものを向上させる技術が必要である。前者については、アノテーションツールによる内容記述処理の自動化によってある程度は達成されているが、オブジェクト認識・トラッキング等の画像処理や発話内容に基づく談話構造化など、改良の余地はまだ多く残されている。後者については、発話文認識精度向上のための言語モデルの修正や、認識結果に対する事後処理としての認識誤り訂正処理が挙げられる。

特に、入力音声については、話題に応じた言語制約を加えることにより精度改善が期待される。音声対話システムの場合はリアルタイム性を重視するために、通常音声認識処理は 1 度しか行われぬが、本研究のようなオフライン処理の場合は、音声認識と言語モデル修正を繰り返し行うことにより、認識結果の精度改善を図ることが可能であると考えられる(図 7)。言語モデルの修正には、認識結果以外に、クローズド・キャプション、Web テキスト [13]、アノテーション情報等の外部知識が利用できる。

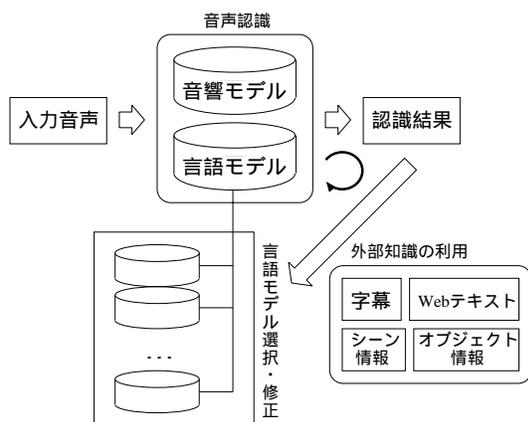


図 7: 話題に応じた言語モデルの選択・修正処理例

また、内容記述量・精度を客観的に示す評価尺度が必要である。コンテンツ時間長に対する内容記述

量や、発話文認識結果の認識精度、構文としての正しさ、等を統一的な評価尺度で測った上で検索時にその評価量を導入することにより、検索結果に対する信頼性を向上させることが可能であると考えられる。

## 5.2 内容記述コスト

アノテーション付与において内容記述にかかるコストは、本システムにおける最大の課題である。内容記述にともなう機械処理の精度を極限もしくはタスク達成に必要な精度まで向上させることが最も重要であることは言うまでもないが、現実的に実用レベルのアノテーションを付与するためには人間の介在が不可欠であるため、人間がスムーズに作業を行うための入力支援やインターフェースの改良も必要である。

現時点の多言語ビデオトランスクリプターを用いて、Windows PC (CPU: PentiumIII 800MHz, Memory: 512MB) 上でアノテーションを行った場合にかかるコストを表 3 に示す。アノテーション作業は、まずツールの使用方法を説明した後にテストファイルを用いて 10 分間操作に慣れてもらい、その後、新しいビデオファイル (88 秒) に対して作業を行ってもらった。これを計 3 人分計測した。シーンとオブジェクトについては、作業者によって結果が異なるため、作業時間をシーン数、オブジェクト数で平均してある。

表 3: 内容記述コスト

分類	作業者			平均
	A	B	C	
(a)	58			
(b)	88	70	74	77
(c)	23	30	17	23
(d)	528	394	425	449
シーン数	5	4	4	
オブジェクト数	2	3	2	
(b-d)	1014	764	755	844
計	1072	822	813	902

(認識結果の単語正解率 78.5%, 正解精度 73.9%)

- (a) 機械処理 (音声認識, シーン検出) 時間 [sec]
- (b) シーン統合・内容記述時間 [sec]
- (c) オブジェクト切り出し・内容記述時間 [sec]
- (d) 発話内容修正時間 [sec]

表3より、音声認識精度が70~80%の場合、アノテーション作業を行う人間にかかる時間的なコストは、機械処理にかかる時間の約15倍、アノテーションを付与する対象データ時間の10倍程度であることがわかる。誰もが手軽にアノテーションを付与できるようにするためには、今後さらなる改良が必要であると思われる。

## 6 まとめと今後の課題

我々が目指しているマルチメディアコンテンツ検索システムの概要について説明し、アノテーションに基づいたビデオ検索を実現する手法を実際のシステムを例に提案した。また、現状のシステムの問題点として、アノテーションにおける内容記述量・精度と内容記述コストを挙げ、これら解決すべき問題について考察した。

今後は、システムの問題点を改善しながら個々の技術を検討し、大量のデータに対して提案した検索手法の有効性を評価していきたいと考えている。

また、昨年度のTREC-2001よりVideo Retrieval Trackが導入され[14]、ビデオデータにおける類似検索の研究促進が期待されているので、こちらの動向にも注目しながら研究を進めていく予定である。

### 謝辞

本研究の一部は、早稲田大学理工学総合研究センターの研究課題「マルチモーダル情報空間における統合的ヒューマンインタフェースに関する研究」によるものである。ここに記して謝意を表する。

### 文献

- [1] 長尾確, 白井良成, 橋田浩一, "アノテーションに基づく知的マルチメディア処理", 情処研究報告, ICS-120-6, pp.41-48, 2000.
- [2] S. Ohira, M. Yoneoka, K. Nagao, "A Multilingual Video Transcriber and Annotation-based Video Transcoding," Proc. of CBMI'01 Workshop, pp.133-140, 2001.
- [3] K. Nagao, S. Ohira, M. Yoneoka, "Annotation-Based Multimedia Summarization and Translation," Proc. of COLING2002, 2002.
- [4] Katashi Nagao, "Semantic Annotation and Transcoding: Making Web Content More Accessible," IEEE MultiMedia, Vol.8, No.2, pp.69-81, 2001.
- [5] 森靖英, 高橋裕信, 岡隆一, "画像と単語の相互検索手法", 人工知能学会研究会資料, SIG-CII-2000-NOV-17, 2000.
- [6] 西田昌史, 緒方淳, 有木康雄, "話者と発話内容の同時検索に関する検討", 人工知能学会研究会資料, SIG-CII-2000-NOV-12, 2000.
- [7] 森靖英, 岡隆一, 他, "WWW上の文書・画像混在データのクロスメディア検索", 人工知能学会研究会資料, SIG-CII-2001-MAR-06, 2001.
- [8] 西崎博光, 中川聖一, "音声入力によるニュース音声検索システム", 電子情報通信学会技術研究報告, SP99-108, pp.91-96, 1999.
- [9] 岡隆一, 高橋裕信, 西村拓一, 他, "パターン検索のアルゴリズム・マップ-CrossMediatorを支えるもの", 人工知能学会研究会資料, SIG-J-A101, pp.1-6, 2001.
- [10] Google:  
<http://www.google.co.jp/>
- [11] World Wide Web Consortium:  
eXtensible Markup Language (XML).  
<http://www.w3.org/TR/PR-xml-971208>.
- [12] 伊藤克巨, 田中和世, 中沢正幸, 岡隆一, "ニュース音声コーパスの構築", 日本音響学会講演論文集, pp.171-172, 1999.
- [13] 伊藤克巨, 秋葉友良, 藤井敦, "WWWは大語彙連続音声認識の学習データとして使えるか?", 日本音響学会講演論文集, 3-9-1, pp.131-132, 2002.
- [14] TREC-2002 Video Track:  
<http://www-nlpir.nist.gov/projects/trecvid/>