■2 群（画像・音・言語）－11 編（マルチメディア）

2 章 マルチメディアコンテンツ解析

（執筆者：佐藤真一）［2011 年 2 月 受領］

■概要■

本章では，マルチメディアコンテンツ解析技術を中心に，マルチメディア情報の計算機による加工の基本技術について概説する．マルチメディア，特に音響・画像・映像のような情報は，数値・テキスト情報と比べて抽象度の低い情報であり，計算機で扱うのには困難が伴うが，人間にとっては直感的で理解しやすい情報である．これは，データとしてのマルチメディア情報は極めて具体的な情報であり，それらに対応する意味レベルの情報とは，セマンティックギャップと呼ばれる極めて大きな乖離が存在する一方，人間は容易にこれらのマルチメディア情報から意味レベルの情報を解釈することができ，人間の期待とマルチメディアを扱うシステムの振る舞いとの間にもずれが生じてしまうことが一因である．メタデータにより意味レベルの記述をあらかじめ与えてしまうという方法もあるが，メタデータ作成の手間や，記述力の限界などの問題があり，やはりマルチメディア情報から自動的に意味情報を抽出する，コンテンツ解析技術の実現が求められている．

本章ではまず，マルチメディア情報，特に映像情報を例に取り，そこから意味レベルの情報を抽出するためのマルチメディアコンテンツ解析技術について概説する．マルチメディア情報は複合情報であり，本における章や節に対応するような構造が多くみられる．特に映像においては，ショットやシーンなどがこうした構造に対応するが，こうした構造を抽出して利用する技術についても概説する．解析の結果，マルチメディア情報を利用するもっとも有望な技術の一つとして，大量のマルチメディア情報から所望の情報を取り出す検索技術が考えられる．そこで，マルチメディア情報に対する検索技術の動向について解説する．また，コンテンツ解析を踏まえ，マルチメディア情報の利用法として有望な加工技術として，編集・要約・オーサリング技術について述べる．最後に，コンテンツ解析の性能の客観的な評価に極めて重要なベンチマークについて紹介する．

【本章の構成】

本章では，自動アノテーション・マルチメディア意味解析（2-1 節），映像構造化（2-2 節），検索（2-3 節），編集・要約・オーサリング（2-4 節），ベンチマーク（2-5 節）について述べる．

■2 群 － 11 編 － 2 章

2-1 自動アノテーション・マルチメディア意味解析

（執筆者：新田直子）［2010 年 2 月 受領］

マルチメディアコンテンツを有効利用するためには，ユーザの要求に対して容易に視聴，検索，編集などができるようマルチメディアコンテンツに対して何らかの意味情報をもたせる必要がある．これを実現する方法の一つとして，マルチメディアコンテンツ内の時区間（セグメント）の意味内容を解析し，その記述をアノテーションとして付与する技術が多く提案されている．一般的に，記述される意味内容の基本としては，いつ（WHEN），どこで（WHERE），だれが（WHO），どのように（HOW），なぜ（WHY），何をした（WHAT）といった 5W1H に関する情報が想定される．マルチメディアコンテンツとは，画像，音，言語といった複数の情報ストリームから構成されるコンテンツ全般を意味するが，本節では，代表的なマルチメディアコンテンツとしてテレビで放送される放送型映像（以下，単に映像と呼ぶ）を対象に，マルチメディア意味解析，自動アノテーションを実現する様々なアプローチについて説明する．

2-1-1 ルールに基づくアプローチ

映像にはニュース，スポーツ，ドラマなど様々なジャンルがあり，各ジャンルには様々な番組が属するが，特定のジャンルや番組において，同じ意味内容の時区間が類似した特徴をもつ場合がある．以下ではまず，このような特徴に関して予め設定したルールに基づいて解析を行うアプローチを紹介する．

(1) 画像に基づくアプローチ

映像における画像ストリームは最下層の一枚一枚の画像であるフレーム，同じカメラで撮影された連続したフレーム列であるショット，意味的なまとまりをもつ連続したショット列であるシーンというように階層化することができ，画像ストリームはシーンの連なりで構成される．このシーン列は，特定ジャンルや番組の映像において，ある定まった構造をもつ場合がある．例えば，ニュース映像は複数のニュースにより構成され，各ニュースはアナウンサによるニュースの紹介シーンと紹介されたニュース自体のシーンの並びにより構成される．スポーツ映像は複数のプレイにより構成され，各プレイは，野球では投球シーン，テニスではサーブシーンなど，視覚的に非常に類似したシーンから始まる．そこで，このような画像ストリームにおける特定パターンを各ジャンルのシーン構成ルールとして予め設定することによる，ニュース映像に対する各ニュースへの分割 [1] や，スポーツ映像に対する各プレイへの分割 [2] など，画像ストリームを利用した，意味をもつ映像セグメントへの分割手法が多く提案されている．また，例えば野球映像においてホームランのシーンは，投球，外野へのカメラのパン，ホームベースへの走塁，などのシーンの並びにより構成されるなど，特定の意味内容に対してシーン構成が定まることもあり，分割した映像セグメントに対し，“ホームラン”といった意味内容記述を付与することも可能である．

(2) 画像・音・テキストの協調的アプローチ

人間が映像を視聴する際，映像を構成する複数のストリームから複合して情報を取り出すことを考えると，(1)項で紹介した手法のように画像ストリームのみからニュース内容や打者名など各映像セグメントの詳細な意味を理解することは困難であり，現実的ではない．よって，映像の意味内容は，画像ストリームの他に，音声，音楽，効果音，雑音などの様々な音源の混合により構成される音ストリーム，画像に存在するテロップ，音ストリームの写しであるクローズドキャプション（CC）などの言語ストリームなど，複数の情報ストリームの意味的関連性に着目した協調解析であるインターモーダル協調 [3, 4, 5] により抽出することが望ましいと考えられる．実際，映像の意味内容解析の関連研究にはインターモーダル協調を用いたものが多く見られる．例として，スポーツ映像に対する，CC からのキーワード，音ストリームからのボリューム上昇，画像ストリームからの色情報の抽出による得点プレイシーン抽出 [4] や，CC からのキーフレーズ抽出と画像ストリームを用いたプレイシーン抽出の統合によるプレイシーンに対するプレイ名，選手名抽出 [5]，また，ニュース映像に対する，CC 及びテロップ中の人物の名前と画像ストリームから抽出された顔領域の共起関係に基づいた人物の同定 [6] などの研究が報告されている．ここで，これらの研究で用いられている CC の代わりに音ストリームに対する音声認識結果を利用することも可能である．音声認識は，ニュース映像では，アナウンサの発話が書き言葉かつ明確であるため，自動字幕作成などにも用いられるほど性能が高い．一方，スポーツ映像のように歓声などの背景音を多く含む話し言葉による発話については認識が困難となるため，CC の利用が望ましいと考えられる．

(3) 映像とメタデータの協調的アプローチ

(2)項で紹介した研究は映像そのものを構成する情報ストリームを用いているが，ほかに，シナリオや電子番組表（EPG），ウェブテキストといったメタデータと呼ばれる映像データに関するデータも同様に重要な情報源として協調的な解析に用いられる．ただし，これらのメタデータは映像と独立して作成されるため，解析に利用する際には，映像との同期付けが不可欠となる．このような手法の例として，スポーツ映像に対するウェブ上の試合結果情報を用いたプレイシーン検出法 [7] がある．この手法では，ウェブ上の試合結果情報にプレイ名や選手名と共に記述される各プレイの試合における発生時刻を，画面上で試合の進行状況を伝えるテロップから文字認識により抽出した時刻情報と対応付けることにより各プレイシーンを抽出し，該当時刻のプレイ名や選手名をアノテーションとして付与している．ウェブ上の試合結果情報は CC に比べ，試合情報が簡潔に記述されており，解析が比較的容易となる．テロップ認識については，テロップの抽出と文字認識が必要となるが，同一の放送局や番組などに対しては決まった色，形のテロップ，文字が一定の位置に出現するため，予め設定したモデルとの比較など簡易な方法により実現可能である．また，シナリオを用いた映画における登場人物の同定手法も提案されている [8]．映画においては，ニュースやスポーツ映像と異なり, シーンと同期した音ストリームにおいて登場人物の名前が発話されることが少ない．そこで，発話者の名前が記述されたシナリオを用い，シナリオにおける発話内容と CC における発話内容の対応付けにより発話者の同定を実現している．

(2), (3)項のように, インターモーダル協調においては, 様々な情報の利用が考えられるが,

共通して，画像処理に必要な計算量を減らし，効率的に信頼性の高い意味内容解析を実現できるという利点がある．特に，人手で作成されたメタデータが入手可能な映像に対しては，効率的に詳細な意味内容解析が実現される．

2-1-2 コーパスに基づくアプローチ

2-1-2 項で紹介した手法では, 抽出したい意味内容に関連した各情報ストリームにおける特定パターンを各ジャンルや番組映像のルールとして予め人手で設定する必要や，各映像に対して人手で作成したメタデータを入手する必要があった．そこで，人間の負担を減らし，かつ汎用性を向上させるため，各情報ストリームにおける特定パターンを，意味内容の正解データ付きの大規模コーパスから統計モデルにより自動的に学習し，内容が未知の映像に対し，統計モデルへの適応度に基づき，セグメントの分割やセグメントの意味内容を決定するアプローチが盛んに取り組まれている．映像における特定パターンの学習に頻繁に利用される統計モデルとして隠れマルコフモデル（Hidden Markov Model）があり，例えば，サッカー映像に対する，プレイシーン及びそれ以外のシーンにおける色や動きに関する特徴の変化パターンの学習に基づいたシーン分割に用いられている [9]．また，野球映像に対する，ホームラン，外野フライ，内野フライ，などの各プレイにおける色やエッジ，歓声，拍手などの音の種類, キーワードの有無といった特徴の最大エントロピーモデル（Maximum Entropy Model）による学習に基づいたプレイ抽出法も提案されている [10]．このようなコーパスに基づくアプローチは，映像の意味内容解析において非常に重要なアプローチであるが，使用する統計モデルや各情報ストリームから抽出する特徴量が性能を大きく左右する．これらを評価する問題としては，共通テストコレクションを用いたビデオ検索技術に関する研究開発促進のための国際ワークショップ TRECVID [11] において，教室，橋，犬，台所，バス，港，夜景など，様々な意味内容のシーンを抽出する高次特徴抽出タスクも設定されている．

2-1-3 むすび

本節で紹介した解析手法により得られる意味内容記述のほとんどが 5W1H のうち, WHO, WHAT，WHEN，WHERE に関するものであり，WHY や HOW についての自動獲得については，サッカー映像に対する選手とボールの軌跡に基づいた戦術解析 [12] などがあるものの，これからの研究課題である. また, これらの記述は自動的に得られたメタデータと考えられる．マルチメディアコンテンツに関するメタデータの記述方式としては，MPEG-7（Multimedia Content Description Interface）が国際標準化されており，主に画像や音などの信号から簡単に抽出される色，形，音量といった低レベルな情報から，5W1H のような高レベルな意味内容情報まで幅広い情報の記述が可能である．今後，自動解析により得られた意味内容をもとにメタデータを自動的に生成することにより，意味内容に基づいた検索や編集などのアプリケーションへの応用が期待される．

■参考文献

1) H.-J. Zhang, S. Y. Tan, S.W. Smoliar, and G. Yihong, “Automatic Parsing and Indexing of News Video,” Multimedia Systems, vol.2, no.6, pp.256.266, 1995.
2) D. Zhong and S.-F. Chang, “Structure Analysis of Sports Video Using Domain Models,” Proc. IEEE International Conference on Multimedia and Expo, pp.920.923, 2001.

3) N.Babaguchi, Y.Kawai, and T.Kitahashi, “Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration,” IEEE Transactions on Multimedia, vol.4, no.1, pp.68-75, 2002.
4) 宮内進吾, 馬場口登, 北橋忠宏, “テキスト・音声・画像の協調的処理による放送型スポーツ映像におけるハイライト検出とインデクシング,” 電子情報通信学会論文誌, vol.J85-D-II, no.11, pp.1692-1700, 2002.
5) N. Nitta, N. Babaguchi, T. Kitahashi, “Generating Semantic Descriptions of Broadcasted Sports Videos Based on Structures of Sports Games and TV Programs,” Multimedia Tools and Applications, vol.25, no.1, pp.59-83, 2005.
6) 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.
7) N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, “Personalized Abstraction of Broadcasted American Football Video by Highlight Selection,” IEEE Transactions on Multimedia, vol.6, no.4, pp.575.586, 2004.
8) R. Turetsky and N. Dimitrova, “Screenplay Alignment for Closed-System Speaker Identification and Analysis of Feature Films,” Proc. IEEE International Conference on Multimedia and Expo, vol.3, pp.1659.1662, 2004.
9) L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, “Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models,” Pattern Recognition Letters, vol.25, no.7, pp.767.775, 2004.
10) Y. Gong, M. Han, W. Hua, and W. Xu, “Maximum Entropy Model-based Baseball Highlight Detection and Classification,” Journal of Computer Vision and Image Understanding, vol.96, no.2, pp.181.199, 2004.
11) http://www-nlpir.nist.gov/projects/trecvid/
12) G. Zhu, Q. Huang, C. Xu, Y. Rui, S. Jiang, W. Gao, and H. Yao, “Trajectory Based Event Tactics Analysis in Broadcast Sports Video,” Proc. ACM Multimedia, pp.58.67, 2007.

■2 群 － 11 編 － 2 章

2-2 映像構造化

（執筆者：井手一郎）［2011 年 3 月 受領］

2-2-1 映像の構造化

映像には，未編集映像（編集前の素材映像，監視映像，ホームビデオ映像など）と編集映像（放送映像，映画など）がある．

編集映像の構造化とは，どのような映像に対しても一律に信号処理的に解析できる低次の構造（以下，「物理構造」）に加え，撮影者・編集者の意図に基づく高次の構造（以下，「意味構造」）を抽出することである．

一方，未編集映像の構造化とは，撮影対象や状況をモデル化したうえで，イベントの発生やカメラワークの検出に基づいて意味構造を抽出することである．

一般に構造化の対象になるのは，複数の未編集映像から編集された映像であるため，本節では主として編集映像，特に放送映像の構造化について記す．

(1) 映像の構造

映像の構造には，与えられた 1 本の映像内の構造と，複数の映像間の構造が考えられる．

(a) 映像内の構造

図 2･1 に映像内の物理構造と用語を示す．

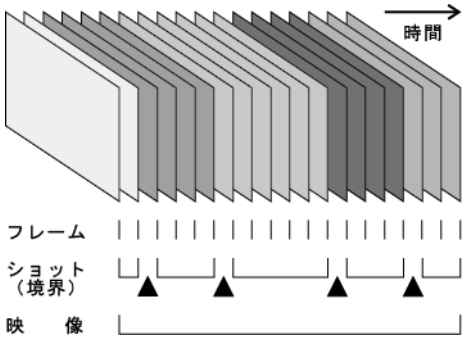

図 2・1　映像内の物理構造

映像の表示は，人間の視覚の残像効果を利用して，一連の静止画像であるフレームを順次高速に切替えて表示することで，あたかも被写体が動くように見せかけることにより実現される．この原理の起源は，19 世紀末にアメリカの T. Edison が発明したキネトスコープ（Kinetoscope），それを改良してフランスの A. & L. Lumiére 兄弟が発明したシネマトグラフ（Cinématographe）であり，今日にいたるまで変わらない．

ショットとは，連続して撮影された一連のフレーム系列である．ショット境界では撮影が中断されたり，他のカメラに切り替わったりするため，前後のフレームは画像的に不連続である．最も単純なショット境界はカット（Cut）と呼ばれ，単純に前後の映像を結合したものである．一方，映像編集技術の発展に伴い，ワイプ（Wipe），フェードイン・アウト（Fade-in,

-out），ディゾルブ（Dissolve）といった高度なショット境界演出技法が登場し，特に放送映像で多用されている．ショット境界の検出技術について，特にカット検出については，古くからフレーム間の不連続点を検出する様々な手法が研究され [1]，既に実用化の域に達している．他の高度な演出技法の検出技術については，様々な手法が提案されているが [2, 3]，演出技法自体が日々発展していることもあり，依然課題が残る．なお，ショット境界検出は，参加型映像検索ワークショップ TRECVID〔本章 2･5 節参照〕において，2001 年から 2007 年までタスクとして存在した．

一方，映像内の意味構造は，本来制作者が何らかの意図をもって編集した構造であり，対象映像ジャンルの習慣や方針に応じて，様々な構造が考えられる．例えば，映画やドラマ，ドキュメンタリといったジャンルの映像は，映像文法 [4] と呼ばれる様々な経験則を基礎として製作される．また，ニュース映像の一般的な意味構造として，キャスタショット（アンカショット，スタジオショットなどとも呼ばれる）で始まり，様々な素材映像や中継映像を結合した複数のショットからなるニュースストーリーを単位とすることが多い．

一方，特定の構図や役割をもつシーン（Scene）に注目して，意味構造の解析に利用することもある．例えば，ニュース映像の「インタビューシーン」，スポーツ映像の「得点シーン」，料理番組映像の「手元シーン」というように，対象映像ジャンルごとに重視する内容に則した性質の映像に注目して検出・利用することが多い．シーンはショット単位で扱うことが多いが，用途によってショットの上位構造や下位構造，更にショットとは無関係な区間と定義されることもあるため，一般的な定義はない．

(b) 映像間の構造

映像間の構造とは，番組など映像の単位をまたぐ何らかの関連に基づく構造のことである．ここでも物理構造と意味構造に分けて考える．

映像間の物理構造は，異なる映像に共起する低次特徴を関連付けた構造である．近年特に注目されているのは，同一映像や準同一（Near-duplicate）映像の共起に基づく物理構造である．

この構造は，特に放送映像において，特定の映像が番組構造を表すのに用いられたり，同一素材映像が若干の編集を経て配信・再利用されるという性質に基づき，意味構造を推定するために利用されることが多い．

一方，意味構造は，映像の意味内容に応じて異なる映像同士を関連付けた構造である．様々な低次特徴から得られる物理構造を組み合わせることにより抽出されることが多い．この際，画像情報だけでなく，付随するテキストや音声の情報による物理構造を利用することが極めて有効である．

映像間の構造化については，特にニュース映像において，異なる日に放送されたニュースストーリーを関連付ける研究が盛んである．

(2) 構造化の方法

映像の構造化の際に，映像特徴のみにより構造化する方法と，映像と強く関連する付随情報を利用する方法がある．

(a) 映像特徴による構造化

一般に，映像は動画像だけでなく，音声，同期して提供されるテキスト情報などからなる．

映像特徴による構造化は，これらをマルチメディア情報として統合的に解析する．

(b) 関連情報による構造化

映像には様々な形態で関連情報が存在することがある．映像と直接関連する情報として，宣伝映像，台本，副読本などがあり，対象映像ジャンルに応じた構造化が考えられる．また，直接同期して提供される情報でなくとも，スポーツのルールのような対象映像ジャンル固有の知識や，関連する内容について記された新聞などの紙媒体やウェブ上のテキスト情報などの利用も考えられる．

2-2-2 種々の映像における構造化

信号処理的に一律に解析できる物理構造に対して，映像の意味構造を解析する際には対象映像ごとの性質や関連情報の有無，更に構造化の用途に特化した処理が必要である．以下にいくつかの映像ジャンルにおける構造化の実例を紹介する．

(1) ニュース映像における構造化

ニュース映像の構造化は，Informedia プロジェクト [5] における様々な試みを始め，映像処理分野において長らく取り組まれている課題である．

ニュース映像の構造化は，番組内の構造解析と番組間にまたがる構造解析に分けて考えられる．前者は一般に，事前処理としてオープニング，天気予報や市場情報といった定型的な映像を除去した後に，各ニュースストーリーを分割する．このような構造化により，1 本の番組の映像をニュースストーリー単位で検索・閲覧することができるようになる．多くのニュース番組において，新しいニュースストーリーが始まるたびに，スタジオにいるキャスタ（アンカパーソン）の映像が使われる傾向が高いことから，キャスタショットを手がかりに分割することが多い．しかし，映像表現技術が発展するにつれ，キャスタショットの検出自体が困難になりつつあることや，必ずしもキャスタショットの出現と新規ストーリー開始が共起しないことから，他の画像特徴のほか，テキストや音声の特徴を併用するのが確実である．

一方後者は，番組内の構造に基づいて抽出されたニュースストーリーについて，番組間にまたがった関連付けを行う．構造化方法としては，ニュースという対象を考慮して，時間（When），場所（Where），人物（Who），トピック（What）といったいわゆる 5W1H の視点から，様々な試みがなされている．例えば，Christel らは映像をニュースストーリーの発生地点に対応付け，現実世界の地理情報を手がかりにして映像間の構造化 [6] を実現した．一方，Satoh らは映像に出現する人物の名寄せ・顔寄せにより人物を手がかりにした構造化 [7] を実現した．更に井手らは，ニュースの時間的変遷を反映したニューストピックの時系列意味構造を解析する手法 [8] を提案した．

(2) スポーツ映像における構造化

スポーツ映像の構造化は，主にハイライトシーン検出やスコア表作成，戦略分析を目標として行われる．

比較的長時間であるというスポーツ映像の特徴をふまえ，ハイライトシーンに基づく構造化は，要約映像の作成などのために必要である．実際に，喚声，特徴的な構図，編集効果の

検出[9] など，様々なハイライトシーン検出手法が提案されており，家電製品へ組み込まれた例[10] もある．

一方，上記の目的に加え，スコア表の自動作成や試合の戦略分析のためには，より詳細な構造化が必要である．一般にスポーツ（特に球技）は定型的な試合構造をもつ．そのため，スポーツ映像を構造化する際には，何らかの形で試合構造を考慮することが有効である．このような考え方に基づいて新田らは，スポーツの試合構造とスポーツ番組の編集方法を考慮して，フットボール映像と野球映像を構造化した[11, 12]．また，Delakis らも同様に試合構造と編集方法を考慮して，隠れ Markov モデルにより，テニス映像を構造化した[13]．これらの手法は，特徴的な音声の検出，クローズドキャプション中の特定の語（選手名，プレー名など）の出現，特徴的な構図，編集効果（特にリプレイ）の検出など，映像に含まれるマルチメディア情報を統合的に処理することにより，構造化を実現している．

(3) その他の映像における構造化

一般の映像や，その他の映像ジャンルについても，様々な構造化の試みが行われている．一般の放送映像群を対象とした構造化として，佐藤らは大規模な放送映像コーパスについて，様々な映像特徴を抽出し可視化するとともに，共通する構造を抽出する手法を提案した[14]．また，柴田は放送映像の内容記述モデルを提案し，ドキュメンタリ映像を例題として，階層的意味構造の構築手法を提案した[15]．

一方，外部の関連情報を利用した様々な構造化が試みられている．例えばドラマ映像を対象として，柳沼らは台本と対応付けることで，台詞構造に合わせて映像を構造化した[16]．森山らは，ショットや音響効果の構造に注目して，心理的印象に基づいて映像を構造化した[17]．三浦らは，料理映像をレシピの調理手順構造と対応付けることで構造化した[18]．他に，放送映像以外の映像として，講義映像を講演スライドと対応付けることで構造化する研究[19] もある．

■参考文献

1) 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, Apr. 1992.
2) 谷口行信，外村佳伸，浜口　洋，“映像ショット切替え検出法とそのアクセスインタフェースへの応用，” 信学論(D-II), vol.J79-D-II, no.4, pp.538-546, Apr. 1996.
3) 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.J91-D, no.10, pp.2529-2539, Oct. 2008.
4) J. Monaco, “How to read a film,” Oxford University Press, Oxford, UK, 1977.
5) 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, Sep. 1996.
6) M. G. Christel, A. G. Hauptmann, H. D. Wactler, and T. D. Ng, “Collages as dynamic summaries for news video,” Proc. 10th ACM International Multimedia Conference, pp.561-569, Dec. 2002.
7) 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, Jan.-Mar. 1999.
8) 井手一郎，木下智義，高橋友和，孟　洋，片山紀生，佐藤真一，村瀬　洋，“大量ニュース映像を対象とした時系列意味構造に基づく情報編纂手法の提案，” 人工知能学論，vol.23, no.5, pp.282-292, Sep. 2008.
9) N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, “Personalized abstraction of broadcasted American football video by highlight selection,” IEEE Trans. Multimedia, vol.6, no.4, pp.575-586, Aug. 2004.

10) I. Otsuka, K. Nakae, A. Divakaran, K. Hatanaka, and M. Ogawa, “A highlight scene detection and video summarization system using audio feature for a personal video recorder,” IEEE Trans. Consumer Electronics, vol.51, no.1, pp.112-116, Feb. 2005.
11) 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-II, no.8, pp.1838-1847, Aug. 2001.
12) 新田直子, 馬場口登, “放送型スポーツ映像の意味内容獲得のためのストーリー分割法,” 信学論(D-II), vol.J86-D-II, no.8, pp.1222-1233, Aug. 2003.
13) M. Delakis, G. Gravier, and P. Gros, “Multimodal segmental-based modeling of tennis video broadcasts,” Proc. 2005 IEEE International Conference on Multimedia and Expo, pp.546-549, Jul. 2005.
14) 佐藤　隆, 児島治彦, 阿久津明人, 外村佳伸, “映像コーパスの構築と分析,” 信学論(D-II), vol.J82-D-II, no.10, pp.1552-1560, Oct. 1999.
15) 柴田正啓, “映像の内容記述モデルとその映像構造化への応用,” 信学論(D-II), vol.J78-D-II, no.5, pp.754-764, May 1995.
16) 柳沼良知, 坂内正夫, “DP マッチングを用いたドラマ映像・音声・シナリオ文書の対応付け手法の一提案,” 信学論(D-II), vol.J79-D-II, no.5, pp.747-755, May 1996.
17) 森山　毅, 坂内正夫, “ドラマ映像の心理的内容に基づいた要約映像の生成,” 信学論(D-II), vol.J84-D-II, no.6, pp.1122-1131, Jun. 2001.
18) 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の対応付け,” 信学論(D-II), vol.J86-D-II, no.11, pp.1647-1656, Nov. 2003.
19) F. Wang, C.-W. Ngo, and T.-C. Pong, “Structuring low-quality videotaped lectures for cross-reference browsing by video text analysis,” Pattern Recognition, vol.41, no.10, pp.3257-3269, Oct. 2008.

■2 群 － 11 編 － 2 章

2-3 検　索

（執筆者：篠田浩一）［2009 年 7 月 受領］

2-3-1 マルチメディア検索

インターネット，テレビ，ホームビデオなどで，マルチメディアコンテンツが急激に増加している．これらを効率良く検索できる仕組みが求められている．現状では，これらのコンテンツに付与されているラベルは，不十分であり，また不正確なことも多い．そこで，マルチメディアコンテンツを構成する，音，画像，映像などのメディアから特徴を抽出してパターン認識を行うことにより検索を行う，マルチメディア検索技術が盛んに研究されている．

マルチメディア検索では，テキスト検索と同様に，検索クエリーをテキスト（すなわちキーワード）で与える場合と，その他のモードで与える場合がある．後者の例としては，ハミングによる楽曲検索，類似画像（動画）を検索するコピー検出，スケッチを用いた画像検索などがある．ここでは主にキーワードを検索クエリーとする画像・映像の検索を中心に述べる．

テキスト検索はすでに商用で使用されているが，マルチメディア検索はまだ性能が低く，実用段階に至っていない．その主な理由として二つあげられる．一つ目は必要な計算資源がテキスト検索に比べはるかに大きい点である．この問題は昨今の計算機技術の進歩により徐々に解決されつつある．もう一つは，テキスト検索の場合と違い，ユーザのキーワードに対応するマルチメディアコンテンツ中の特徴が何であるかが必ずしも明確ではない，という点である．ユーザが入力するキーワードは一般に何らかの概念を表す（例：自動車，南の島，賑やかな週末など）．これらは，コンセプト，あるいは，高レベル特徴と呼ばれる．一方，マルチメディアから抽出される特徴（例：色，テクスチャ，形状など）は，低レベル特徴と呼ばれる．これら二つの特徴の間には直接的な関係はなく，また，それらを関係づける方法も明確ではない．この乖離は，セマンティックギャップと呼ばれ，その解消は，高性能なマルチメディア検索の実現への大きな課題となっている．

2-3-2 コンテンツ検索

ここでは，「自動車」，「人間」，「椅子」など，画像特徴が比較的明確であり，セマンティックギャップの影響が少ないオブジェクトを対象とした，低レベル特徴を用いた検索について述べる．

図 2・2　コンテンツ検索の対象となるオブジェクトの例

検索処理は，特徴抽出とオブジェクト検出の 2 段階に分かれる．特徴量としては，回転，拡大・縮小，平行移動などに対して不変であり，また，人間が見て重要と思われる特徴を効率的に表現しているものが望まれる．大局的特徴量と局所的特徴量に大別される．

(1) 大局的特徴量

大局的特徴量は画像全体に渡る特徴を表す. 色（Color）, テクスチャ（Texture）, 形状（Shape）を表す特徴量に大別される．以下，色とテクスチャについて代表的な特徴量を示す.

色（Color） 色ヒストグラムは，画面全体の色情報を低次元に圧縮する．例えばコロンビア大学では，HSV 色空間を用い，H 18 次元，S 3 次元，V 3 次元に Gray スケールの 4 次元を加えた 166 次元の特徴量を用いている．Grid Color Moments は，色の位置情報を表す．例えば，3 色（RGB）画像を 5×5 程度に分割し，分割後の各サブ画像において 3 次までのモーメント（平均，分散，歪度）を求める．この場合は 3×5×5×3 で 225 次元である．

テクスチャ（Texture） 単純に画像のフーリエ変換が用いられる．また，位置情報を取り入れた特徴として，フーリエ関数にガウス関数を乗じた Gabor 特徴量がしばしば用いられる．Gabor 特徴量はスケールと方向の 2 次元の特徴で表される．

(2) 局所的特徴量

局所的特徴を求める際には，まず特徴的な領域を検出し（特徴点検出），その後，それを記述する（特徴表現）という 2 段階の処理が行われる．重要点検出においては，その対象となる Interest Point（重要点）をどのように定義するか，がポイントになる．点状の特徴を検出する検出器として Harris-corner Detector, 領域状の特徴を検出する検出器として Hessian-Affine Detector が用いられることが多い.

局所特徴の記述としては，物体の移動や回転，すなわち，アファイン変換に対し不変な記述が望まれる．そのような記述法として，SIFT（Scale Invariant Feature Transform）がよく用いられる．SIFT では，まず検出された領域全体における特徴変化の方向と大きさを計算し，アファイン変換に対し不変となるよう正規化を行う．次に，領域を 4×4 に分割し，分割後の各々の領域において，8 方向における特徴変化の大きさを計算する．この場合，特徴量は 4×4×8 の 128 次元となる.

局所的特徴を用いる場合には，その個数がサンプルにより異なり，またお互いの順序関係がないため，そのまま特徴ベクトルとして表現することは難しい．そこで，テキスト検索の場合と同様，Bag of Word（BoW）のアプローチがしばしば用いられる．BoW アプローチでは，まず，局所特徴量の空間において，局所特徴量の量子化を行う．量子化には k-means アルゴリズムなどのボトムアップクラスタリング手法が用いられる．その結果作成されたコードブックを Visual Dictionary と呼び，その要素を Visual Word と呼ぶ．個々の画像について，位置情報は無視して，各 Visual Word の出現回数をカウントし，それを並べたベクトル（次元数はコードブックサイズに等しい）を，画像の特徴ベクトルとする．テキスト検索の場合と同様に，tf・idf を用いて，検索における重要度を用いて Visual Word に対し重みづけを行ったり，pLSA（Probabilistic Latent Semantic Analysis）や LDA（Latent Dirichlet Allcation）などの手法を用いて次元圧縮を行う処理が行われることもある.

2-3-3 検出器

オブジェクトの検出には，サポートベクターマシン，最大エントロピーモデルなどの識別モデルが用いられることが多い．オブジェクトが含まれる画像を正例，含まれない画像を負例として 2 値判別を行うモデルを学習する．混合ガウス分布などの生成モデル，Adaboost などの集合学習（Ensemble Learning）が用いられることもある．

また，映像の検索においては，計算量削減のため，まず映像をショットごとに分割し，次に各ショットにおける代表的なフレーム（キーフレーム）をいくつか抜き出し，それらのフレームの静止画像に対し，上述の処理を行うことが多い．この場合，必ずしも映像中のすべてのフレームに対象となるオブジェクトが出現しているとは限らないことに注意が必要である．

また，映像コンテンツには音声が付随していることが多く，その場合，音声情報の検索も有用である．特に，ニュース番組など，音声情報が重要な役割をもつコンテンツの検索においては，むしろ音声情報を用いた検索が，画像情報を用いた検索よりも効果的であることがある．もちろん，両者を併用した場合には，それぞれ単独の場合よりも性能が高くことが多い．

2-3-4 概念に基づく検索

前節で述べた低レベル特徴に基づく検索では，ユーザがクエリーとして適当なキーワードを探すのが難しい．いつでもキーワードに対応する意味（概念）をもつオブジェクトがあるわけではない．また，すべての可能なキーワードに対し検出器を用意するのは非現実的である．そこで，実用的なマルチメディア検索の実現には，概念語彙セットの設計，及び，それに基づくクエリー拡張（Query Expansion）が重要となる．

マルチメディア検索における概念語彙は，WordNet などのテキスト検索に用いられる語彙とは異なる．画像から検出が可能で，かつ，データ中の出現回数が多いものが良い．しかし，大部分のデータに出現するものはかえって不都合である．また，定義が明確であり，複数の意味をもったりしないもの，いろいろな用途で使用可能（一般的なもの）なものが望まれる．語彙サイズとしては最低 3000 から 4000 語が必要であると考えられている．これまで開発された語彙セットとしては，TRECVID（20～40 語），MediaMill（101 語），LSCOM（500 語）などがある．

また，テキスト検索の場合と同様，ユーザからのフィードバックを利用したクエリー拡張の効果が高い．そのための，使いやすいインタフェース，ユーザから獲得する情報の選別，それに基づく検索の高性能化などが盛んに研究されている．

2-3-5 今後の課題

マルチメディア検索の性能向上のための最大の課題はデータの不足である．高性能なシステムの検索のためにはある程度の量のタグ付きデータが必要であるが，その作成コストは大きい．Web 2.0 的アプローチとして，協調的タグ付けの研究が行われている．また，オブジェクトが出現した位置や時刻など，より精細な情報のタグ付けが望まれる．更に，有効な特徴量，特に形状特徴に基づく特徴量の研究が重要と考えられる．

■参考文献

1) R. Datta, D. Joshi, J. Li, and J.-Z. WANG, “Image retrieval: ideas, influences, and trends of the new age,” ACM Computing Surveys, vol.40, no.2, Article 5, 2008.
2) J. Sivic and A. Zisserman, “A video google: A text retrieval approach to object matching in videos,” Proc. ICCV, vol.2, pp.1470-1477, 2003.
3) M. R. Naphade, J. R. Smith, J. Tesic, S. F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, “Large-scale concept ontology for multimedia,” IEEE MultiMedia Magazine, vol.13, no.3, 2006.

■2 群 － 11 編 － 2 章

2-4 編集・要約・オーサリング

（執筆者：尾関基行）［2009 年 4 月 受領］

マルチメディアコンテンツ（主に映像）の編集・要約・オーサリングの自動化・省力化のための要素技術について概説する．**図 2･3** にデータと処理の流れを示す．図中で編集・要約・オーサリングと記した部分だけでなく，その他の円で囲まれた部分も本節の範囲に含まれる．

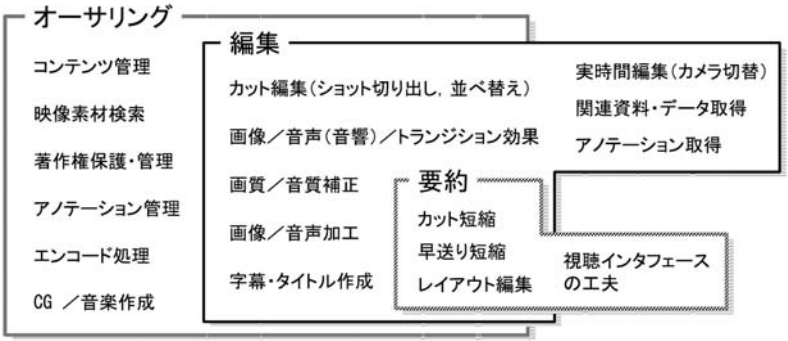

図 2・3　データと処理の流れ

2-4-1 用語について

編集・要約・オーサリングという三つの用語は包含関係にあり，分野によって解釈の幅に違いがある．本節では**図 2･4** のように区分する．編集・要約・オーサリングの主な機能（意味）は，それぞれ次のようにまとめられる．

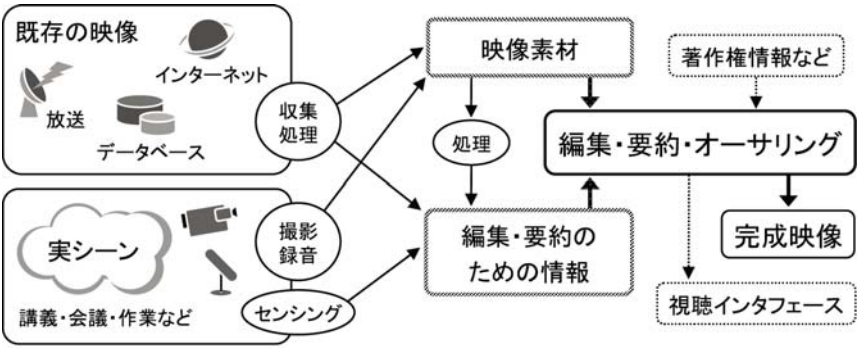

図 2・4　編集・要約・オーサリングの位置づけ

編集： 世界をある視点から写し撮った映像断片を時間軸上に並べて，特定の主旨の下で再構成することをいう．動画と音による世界の時間的・空間的な要約（及び拡張）ともいえる．狭義には，映像素材から必要な部分を切り出し，時間軸に沿って配置すること

を指す（カット編集）．広義には，映像に補正や効果を加え，音楽や字幕，タイトルなどを付与することまで含める．

要約： 視聴者が映像全体の内容を把握するまでの時間を短縮すること，あるいはそのための技術をいう．一般的には映像自体を時間的に縮める処理を指すが，広義には映像のレイアウトや視聴インタフェースの工夫による視聴時間の短縮処理も含む．映像の自動要約は，大量の映像コンテンツを扱う際の必須技術である．

オーサリング： 映像編集とその周辺の処理を合わせた総合的な機能をいう．上述の編集・要約の機能に加え，映像素材の検索やアノテーション管理，著作権保護，エンコード処理などを含む．これらの機能を提供するユーザインタフェースを備えたソフトウェアをオーサリングツールと呼ぶ．一般に，映像編集ソフトウェアとはこのオーサリングツールのことをいう．

2-4-2 編　集

編集を自動化するためには，映像内（あるいは撮影シーン）から編集に利用するための情報を自動抽出する必要がある．このような情報としては，(a)動画の変化量や音のリズムなどの低レベル特徴と，(b)映像中の人物やシーンに関する情報，構図・カメラワークなどの高レベル特徴（概念）がある．低レベル特徴を利用した自動編集には，例えば，与えた音楽のリズムと動画の変化量が合うように編集する手法がある [1]．一方，高レベル特徴を利用した自動編集には，構図やカメラワークの情報を利用し，映画の文法 [2)*1] に沿って自動編集するものがある．例えば，ショットの長さや配置順などをもとに評価関数や制約条件を定め，制約充足問題として自動編集を扱うことができる [3]．

シーンの撮影と同時に編集まで行う実時間編集と呼ばれる技術がある．これは主に遠隔講義や遠隔会議で利用される．実時間編集における典型的な編集手法はカメラ選択（ショット切替）である．複数のカメラで一つのシーン（会議参加者や講師，生徒，資料など）を撮影し，そのシーンで起こったイベント（人物の位置や発話，板書など）に応じてカメラを選択することで映像を切り替える（4-1 節）．会議や講義以外にも，料理や工作などを対象とした自動編集手法がある [4]．この手法では，説明者が視聴者の注目を集めようとする行動（物を掲げて「これは…」というなど）を手の位置と発話内容をもとに認識し，その注目箇所をクロースアップで映しているカメラに切り替える（**図 2･5**）．このように，被写体の動作に基づいた実時間編集では被写体 = 編集者という関係が生じるため，認識手法や編集結果の評価だけでなく，ユーザインタフェースとしても評価しなくてはならない [5]．

図 2・5　机上作業シーン映像の実時間編集（話者の手の位置と発話内容に基づいて最適なカメラを選択）

*1　映画製作などを通して経験的に蓄えられてきた撮影や編集の基本ルール．

2-4-2 要　約

映像に関する事前情報が無い場合，最も軽処理で効果的な要約方法は映像を単に早送りすることである．カット点（カメラの切替点）から短時間ずつ再生する方法も処理は軽いが，あまり良い要約にはならない．映像に関する事前情報があり，処理に時間がかけられる場合は，早送りよりも優れた要約が生成できる．例えばスポーツ映像の場合，得点に関わる部分などのハイライトシーンを自動抽出して繋ぎ合わせることで大要を損ねずに要約することができる [6]．また，ホームビデオなどの編集が不十分な映像の場合，音声的に盛り上がっている部分のみを再生することで無駄な部分を効率よく省くことができる [7]．

また，映像自体を短縮するのではなく，レイアウトや視聴インタフェースを工夫することで閲覧時間を短縮することもできる．最も簡単な方法は，映像を等時間間隔やカット点で切り分けて，その映像サムネイルを空間配置することである．このアプローチは，特定の構造をもった映像要約では特に高い効果を発揮する．例えば，複数のニュース映像から同じ事件を扱った映像断片を集めてそのサムネイルをグラフにすれば，網羅的な視聴に比べて遥かに効率的に事件の全体像を把握することができる（2-2 節）．ただし，映像サムネイルはその中身が把握できるように（部分的に）再生されている必要があるため，やはり映像自体を短縮する要約技術は不可欠である．

2-4-3 オーサリング

映像編集の準備から後処理までの作業を追うことでオーサリングの機能を概説する．まず，映像編集は映像素材を用意することから始まる．自ら撮影した映像でないならば，著作権（利用規約）や不正コピーを調べる必要がある（3-5 節）．主筋の映像の間を繋ぐカット[*2] は過去の映像を再利用することも多いが，データベースが大きくなると映像検索技術が必要となる（4-5 節）．この映像検索では，概念に基づく検索や低レベル特徴量に基づく検索（2-3 節），その他様々な映像の類似度（1-3 節）を用いて実現される．編集作業を終えると，必要な形式にエンコードし，更に著作権情報やアノテーションを付与する．著作権情報の付与については，複製防止や超流通のための電子透かし技術が利用される（3-5 節）．映像の内容に関するアノテーションを付与しておけば検索や視聴の際に便利だが，手作業での付与は労力が大きいため，様々な自動化の研究がなされている（2-1 節）．

■参考文献

1) Xian-Sheng HUA, Lie LU, Hong-Jiang ZHANG, “AVE - Automated Home Video Editing,” Proceedings of the 11th ACM international conderence on Multimedia, pp.490-497, 2003.
2) ダニエルアリホン（岩本憲児，出口丈人訳）, “映画の文法,” 紀伊國屋書店, 1980.
3) 尾形　涼，中村裕一，大田友一，“制約充足と最適化による映像編集モデル,” 信学論，vol.J87-D2, no.12, pp.2221-2230, 2004.
4) 尾関基行，中村裕一，大田友一，“注目喚起行動に基づいた机上作業映像の編集,” 信学論，vol.J88-D2, no.5, pp.844-853, 2005.
5) 尾関基行，中村裕一，大田友一，“話者の注目喚起行動による机上作業映像の自動編集―ユーザインタフェースの側面からの評価―,” 情報科学技術レターズ, pp.269-272, 2004.
6) 熊野雅仁，有木康雄，塚田清志，“野球中継のハイライトシーン実時間配信を目的とした特徴のマイニ

*2 場所の移動や時間の経過を示唆する映像表現のための短い映像．これらは，「夕焼け」や「飛んでいる飛行機」といったキーワードや，「この映像と似ている映像」という形で検索できると便利である．

ングによる PC シーンの自動検出（<小特集>デジタル放送・伝送方式）,” 映像情報メディア学会誌, vol.59, no.1, pp.77-84, 2005.

7) 湯口昌宏, 宮下直也, 日高浩太, 佐藤　隆, “映像ダイジェスト配信システム「チョコパラ TV」の開発,” 第 2 回デジタルコンテンツシンポジウム講演予稿集, 1-4, 2006.

■2 群 － 11 編 － 2 章

2-5 ベンチマーク

（執筆者：帆足啓一郎）［2009 年 3 月 受領］

世界中の研究機関において，様々なマルチメディアコンテンツ解析技術の研究開発が進められている．こうした研究を推し進め，技術の改良を図るためには，マルチメディアコンテンツ解析の有効性を共通の評価基準やデータに基づいて評価することが重要である．しかし，再現可能かつ有用な検証実験を行うためには，大量の実験用コンテンツの収集，収集した実験用コンテンツの情報を表すメタデータの付与，ならびに，解析技術の有効性を測定するための評価尺度の策定などが必要となるが，こうした評価用データは，データの量や著作権などの権利関係を考慮すると，個々の研究者や研究組織が取りまとめるには大きな困難をともなう．

以上の背景により，マルチメディアコンテンツ解析や検索に関連する研究を評価するためのベンチマーク環境の整備が，近年，急速に進められている．ベンチマークの構成要件としては，以下の項目があげられる．

(1) 実験対象コンテンツ
(2) 実験対象コンテンツに対するメタデータ
(3) ベンチマーク実験における共通タスク
(4) タスクを評価するための共通指標

テキスト情報検索の分野においては，こうしたベンチマークを構築する取組は，TREC [1] や NTCIR [2] といった国際ワークショップの枠組みの中で，1990 年代前半から進められているが，マルチメディアコンテンツ解析分野では，取り扱うデータの容量やコンテンツの著作権などの問題があり，取組が遅れていた．しかし，映像検索における TRECVID ワークショップなど，ようやくマルチメディアコンテンツ解析の分野においても，いくつかのベンチマーク環境が確立されつつある．

本節では，マルチメディアコンテンツ解析に関連するベンチマークの代表的な事例として，静止画コンテンツの「Caltech 256」，ならびに映像コンテンツの「TRECVID」についてそれぞれ説明する．

2-5-1 静止画コンテンツ解析のベンチマーク：Caltech 256

Caltech 256 は，米・カリフォルニア工科大学の Griffin らによって構築された，静止画コンテンツ分類のためのデータコレクションである 3)．本コレクションは，256 種類のカテゴリに分類された，30608 個の静止画から構成されている．静止画コンテンツの多くは，Google イメージ検索などを利用し，Web 上から集められている．具体的には，カテゴリ名称を検索クエリーとして得られた Web 画像検索結果に含まれる静止画に対し，各カテゴリに該当するか否かの判断を，人手で行う．その結果，“good”と評価された静止画を，当該カテゴリの中に含める形で，カテゴリごとのデータを整備している．更に，本コレクションには，どのカテゴリにも属さない“Clutters”に該当する静止画も準備されている．**図 2･6** に，Caltech 256 に含まれる静止画のサンプルを示す．

図 2・6　Caltech 256 の静止画サンプル（左から「American flag」,「Butterfly」,「Hamburger」,「Teddy bear」,「Clutters」のカテゴリ内画像）

Caltech 256 では，2003 年にカリフォルニア工科大で構築された Caltech 101 のデータコレクションで指摘されたいくつかの問題点に対する改善策が施されている. 例えば, Caltech 101 では，すべての画像が横長の構図になっている他，原画像を回転させただけの静止画が含まれているため, 各カテゴリへの分類が容易であるという問題点があった. これに対し, Caltech 256 では，横長以外の構図の静止画が多数含まれている他，原画像を加工しただけの静止画は含まれていない．また，Caltech 101 では，所属している静止画の数が極端に少ないカテゴリがあったため，当該カテゴリへの分類器を構築するための学習データが十分に確保できないという課題もあったが，この課題への対策として，Caltech 256 では，最小カテゴリでも 80 個の静止画コンテンツを揃えている.

上記の工夫により，Caltech 256 では，Caltech 101 と比べて，難易度が高い静止画コンテンツ分類実験を行うことが可能になっている．しかし，図 2･6 に示される静止画の例からも明らかな通り，Caltech 256 に含まれる静止画の多くは，対象カテゴリを明確に示すものがほとんどである．そのため，一般的な静止画コンテンツの分類と比較した場合，Caltech 256 において設定されている分類タスクの難易度は低いと考えられる.

2-5-2 映像コンテンツ解析のベンチマーク：TRECVID

映像コンテンツ解析のためのベンチマークとしては, 米･NIST が主催する TRECVID（TREC Video Retrieval Workshop[4)]）において提供されているデータコレクションが，近年，映像分野においてデファクトスタンダードになっている．TRECVID は，当初 TREC の中の一つのサブタスクとして，TREC 2001 と TREC 2002 に含まれていたが，2003 年からは，独立したワークショップとして開催されており，例年，多数の研究機関が参加している[5)].

TRECVID の実験対象コンテンツは主催者が収集しており，毎年のワークショップ参加者に配布している（近年は，参加者が Web からダウンロードする形をとっている）．2006 年の TRECVID までは，米国・中国などのニュース番組を中心とした映像コンテンツが実験用データとして利用されていたが，2007 年以降は，ニュース以外のテレビ番組や，BBC Rushes と呼ばれる，編集前の撮影映像素材など，実験用映像コンテンツのバラエティは拡張している.

TRECVID の開始以降，すべての TRECVID において実施されているタスクは，高次特徴抽出タスク（High-level feature extraction，以下「HLFE タスク」）と検索タスク（Search）の 2 種類である. HLFE タスクでは, 検索対象映像コンテンツの各ショットの中から,「Classroom」,

「Bridge」,「Emergency vehicle」など，課題として与えられた feature が出現するショットを検出するタスクである．本タスクにおける検出精度評価のための参照用データは，各参加組織によって提出された実験結果を対象に，人手で正解を判断するプーリング方式によって構築される．また，評価指標は，情報検索分野において利用されている再現率（Recall）と適合率（Precision）に加え，2007 年以降は InfAP（Inferred Average Precision）という指標 [6] が採用されている．

検索タスクでは，与えられた検索課題に該当するショットを，検索対象映像コンテンツの中から検索することを目的とするタスクである．本タスクでの検索課題は，テキスト文での表記に加え，正解のサンプルとして，静止画や映像（ショット）によって構成されており，HLFE タスクにおける feature と比べ，内容が複雑である（例：“Find shots of a street scene with multiple pedestrians in motion and multiple vehicles in motion somewhere in the shot.”）．また，本タスクでの実験条件としては, Interactive（検索システムとユーザとのインタラクションあり，定められた時間内での検索精度を評価），Manual（検索課題を元に，ユーザが検索クエリーを作成し，システムに入力），及び Fully automatic（システムが検索課題を入力とし，ユーザの介在なく検索を行う）の 3 種類が設定され（**図 2･7** 参照），各条件下において，検索精度が評価される．

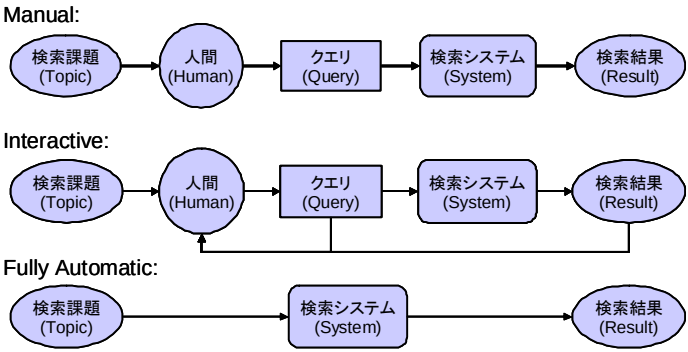

図 2・7　TRECVID・検索タスクの実験条件

上記の HLFE タスクと検索タスクを含め，TRECVID では，毎年のワークショップでの参加者間の議論を経て，その次のワークショップにおける課題が決定される点が特徴的といえる．例えば，初期の TRECVID においては，映像内のカット点検出（Shot Boundary Detection）が共通課題として設定されていたが，参加組織によるカット点検出の精度や処理速度が高止まりしつつあると判断された 2006 年からは，課題として除外されている．一方，2008 年には，新たなタスクとして，監視カメラ映像からのイベント検出（Surveillance Event Detection）や，複製コンテンツ検出（Content-Based Copy Detection）の二つの課題がパイロットタスクとして提案され，TRECVID 2008 での評価を経て，TRECVID 2009 において正式な課題として採用される予定である．

TRECVID のデータコレクションは，映像コンテンツの量も豊富な上，研究目的での映像コンテンツの利用権も許諾されていることから，映像コンテンツ解析技術評価のベンチマー

クとしては理想的なデータの一つといえる．また，評価タスクも，ワークショップ参加者間での協議を経て決定されるため，実世界でのニーズに即した課題が設定されやすいというメリットも大きい．しかし，Caltech 256 と同様，TRECVID において設定されている課題と，実際の一般ユーザが必要とする映像コンテンツ解析技術とは必ずしも一致していないため，TRECVID において高い評価が得られた技術であっても，まだ実用化には至る例は少ないのが現状である．したがって，今後のマルチメディアコンテンツ解析のベンチマークにおいては，一般ユーザのニーズに即した実験用映像コンテンツの準備と課題の設定も，重要な課題と考えられる．

■参考文献

1) Text REtrieval Conference (TREC) Home page, http://trec.nist.gov/
2) NTCIR (NII Test Collection for IR Systems) Project, http://research.nii.ac.jp/ntcir/
3) G. Griffin, AD. Holub, P. Perona, The Caltech-256, Caltech Technical Report, http://www.vision.caltech.edu/Image_Datasets/Caltech256/paper/256.pdf
4) A. Smeaton, P. Over, W. Kraaij, “Evaluation campaigns and TRECVid,” Proceedings of MIR'06, pp.321-330, 2006.
5) 帆足，菅野，松本，“映像情報検索とその評価技術の最前線,” 情報処理学会会誌，vol.46, no.9, pp.1016-1023, 2005.
6) E. Yilmaz, J. A. Aslam, “Estimating average precision with incomplete and imperfect judgments,” Proceedings of ACM CIKM 2006, pp.102-111, 2006.