Keywords you should know. 第29回

知っておきたいキーワード

映像の自動要約技術

(正会員) 滝嶋康弘†

† 株式会社KDDI研究所

"Automated Content Summarization Technology" by Yasuhiro Takishima (KDDI R&D Laboratories Inc., Fujimino)

キーワード：自動要約，メディア解析，コンテンツ，ハイライト，ダイジェスト

映像の自動要約技術とは

昨今のディジタルビデオレコーダやチューナ付きパソコン，さらには携帯電話まで，録画した番組を自動的に要約する機能を搭載する機器が増えています．これらが要約できる映像のジャンルは，**表1**のように多岐にわたります．映像自動要約は，機器のディジタル化，映像処理技術の発達，面白映像制作手法の進展，映像コンテンツ配信手段の多様化，などの背景から，膨大な映像を見ることに多忙を極める現代人には，大変便利なツールとなりつつあります．

しかし，映像を要約するとは，どのようなことなのでしょうか．一言で要約すると言っても，映像コンテンツには，時間が長くさまざまなシーンを含んでいる，ストーリーやシーン変化の流れがあり，時間的な構造を考慮する必要がある，重要あるいは見たい映像がシーンの先頭と同期している保証がない，など独自の課題が多くあります．また，大量の映像コンテンツから，視聴者が見たい番組やシーンを適切に見つけ出すためには，冗長なシーンや意味のないシーンを排除し，盛り上がったシーンやストーリーの把握に必要なシーンだけを抽出できることが重要です．

そのために，自動要約技術では，コンテンツの重要部分を示す手掛かりをディジタル信号から抽出し，短時間で番組コンテンツを視聴できるような仕組みを構築しています．表1のように，厳密にはジャンルごとに要約の目的は多少異なりますが，ある程度共通の要求条件を満たす基本的な処理が有ります．

本稿では，「要約」という処理を，スポーツやニュース，音楽など，盛り上がったシーンや特徴を持つシーンだけを抽出する「ハイライト」と，映画やドラマなど全体のストーリーを保ちながら，重要なシーンを満遍なく見つけ出してくる「ダイジェスト」の二つに大別して（**図1**），一つの処理方式[1)]を例にその内容を紹介します．

表1　映像自動要約が対象とするジャンルとその処理の例

ジャンル	代表的な処理
スポーツ（野球，サッカー，ラグビー，相撲・格闘技，ゴルフ，テニス）	得点や勝敗に関わる，または見ていて面白いなど盛り上がったシーンを抽出
ニュース	記事ごとにアンカーパーソンの出演しているショットを抽出
音　楽	トーク部分を削除，またはトーク部分を抽出
娯楽（競馬，将棋，囲碁）	レースあるいは指し手部分を抽出
ドラマ	会話やアクションなどの代表的なシーンの抽出

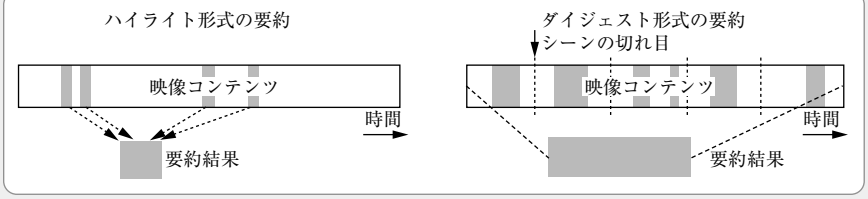

図1　映像コンテンツ要約の形式

映像情報メディア学会誌 Vol. 62, No. 5, pp. 714～716 (2008)

「ハイライト」生成のメカニズム

「ハイライト」形式，「ダイジェスト」形式とも，要約の構成要素は「ショット」と呼ばれる一つのカメラ操作で撮影された一連の画像であり，共通的な処理に基づいて映像コンテンツの中から重要なショットを探し出します．処理の手順を一言で言えば，入力された映像をショット単位に分割した後，各ショットにおける音響特徴および画像特徴を解析することにより，要約区間を適応的に決定します．

ハイライト生成方式では，特にスポーツ番組などで，歓声の大きな部分を手掛かりに，盛り上がりシーンの抽出を行う手法が一般に採られています．さらに，野球の投球シーンやテニスのラリーシーンなど，固定的なショットが明確に存在し，それがイベント（ここでは得点やチャンス・ピンチを引き起こすようなプレーのこと）境界となるスポーツ（以下“構造的スポーツ”）では，イベント境界から始まる歓声周辺の区間を抽出し，一方サッカーなどの固定的なショットを持たないスポーツ（以下“非構造的スポーツ”）では，歓声周辺の区間を重要なイベント（シュートにつながる一連のプレイなど）として抽出します．例えば，野球映像では投球シーンとそれに後続する一連のショットの中で歓声が大きい区間がある場合に，ヒットやホームランなどの重要なイベントが発生したと見なせるわけです．

具体的には，構造的スポーツに対しては，画像特徴を併用してイベント境界を検出します．例えば，画面内の色の空間的な分布を表現する色配置情報に基づいて，頻繁に出現するショット，すなわち，イベント境界を検出することができます．つまり，テレビのスポーツ番組においては，通常カメラの台数やカメラ位置が固定されており，頻出する特徴的なショットを検出することによってイベント境界を特定できるため，これと音響特徴（歓声）との前後関係を利用してハイライトを生成するわけです．音響特徴としては，各ショットにおいて，複数の周波数帯域に分割された音響信号を帯域に応じて重み付けした音響エネルギーとして評価し，そのピークが存在した場合，該当するショットの前にハイライトが存在すると判定します．さらに非構造的スポーツについては，まずイベント境界を決定し，対応するショット以降の一連のショットがピークまたは充分に大きい音響エネルギーを持つ場合，その一連のショットをハイライトとします．**図2**にハイライト生成の概念図を示します．

このような手法で生成されたハイライトの抽出精度ですが，例えば，相撲における取組み，テニスにおけるラリーというように，頻繁に出現するショットそのものが重要なイベントである番組の場合，相撲については全正解のうち95％以上，テニスについては98％を検出することができました．特に，人気力士の取組みやネット際でのプレーなど，注目度の高いイベントは，ほぼ確実に検出できます．なお，生成されたこれらのハイライトは，入力映像の長さのおよそ1/10の長さです．さらに，野球映像において大きな歓声を伴う頻出ショット（投球シーン）を解析したところ，大部分がホームランやヒットなど得点につながるイベントであり，これらも90％程度の高い精度で検出していますし，非構造的スポーツでは，サッカーのシュート・ゴールシーン，ゴルフのカップインのシーン，アメリカンフットボールのタッチダウンのシーンなど，いずれも90％以上の精度で抽出できています．

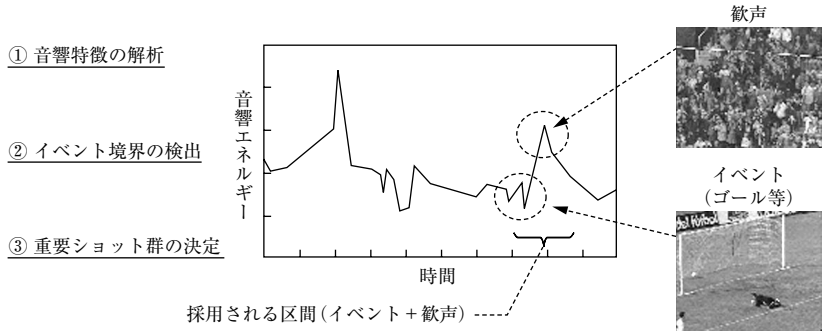

図2　ハイライト生成の概要

「ダイジェスト」生成のメカニズム

ダイジェスト生成は，ハイライト生成に比較して，全体のストーリーを保ちながら，重要なシーンを満遍なく見つけ出してくることが要求されるため，より広域的かつ複雑な処理が必要になります．一つの処理手法[1)]を例にそのメカニズムを説明します．

ダイジェスト生成では，ストーリー性を保つ必要がある一方で，映像コンテンツに依存しない音響特徴および画像特徴を用いることにより，映画やドラマなどのストーリー自体を理解しなくても，予め指定した任意の長さのダイジェストを生成できます．また，音響特徴については信号的な特徴しか利用しないため，言語には依存しない処理が可能です．

具体的には，ダイジェストの構成要素となる重要ショットを決めるため，特に動き特徴に着目します．まず，ある規準に基づいて映像全体を等分割し（等分割された区間つまりショット群を「シーン」と呼ぶことにします），画面内の「動きの強度」を用いて ☞

☞ シーンを「動的」または「静的」に分類します．このとき，「動的シーンではより活発なショットが，静的シーンではより静穏なショットが，それぞれ高い重要度を持つことが多い」という経験則を利用します．例えばアクション映画などでは，興味深いイベントにおいては動的な被写体を多彩なカメラワークで捉えると同時に，さまざまな映像効果を利用して頻繁にショットが切替わる傾向があります．一方ロマンス映画などでは，登場人物の会話ショットが多く出現し，ショット内の動きが少なく平均的なショット長が長いという性質があるわけです．このように動的，静的なシーンに対して，それぞれ特徴的なショットを要約区間と決定していきます．また，ダイジェスト生成においては，音響特徴も非常に重要な手がかりです．これは，映画における銃声・爆発などのイベントや場面に応じたBGMなど，音響情報が重要なシーンに付随することが多いからです．音響特徴としては，ハイライト生成と同様に，例えば複数の周波数帯域に分割された音響信号の，帯域に応じた重み付けエネルギーを用いることができます．ダイジェスト生成処理概要を**図3**に示します．

このようなダイジェスト生成技術によって，映画やテレビドラマからオリジナルの長さの1/10の長さを持つダイジェストを生成した場合，生成されたダイジェストとWebサイトから入手した解説またはあらすじのテキストとの文脈の比較で，70%～80%程度の段落（5分から10分程度の内容が一段落に相当）が，ダイジェストの中で説明されており，文脈を保持したまま要約がされていることがわかります．

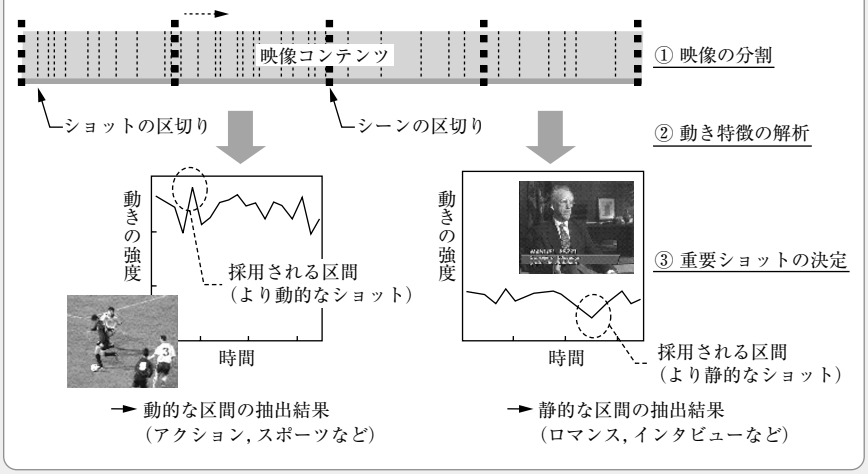

図3　ダイジェスト生成の概要

さらなる展開

本稿で紹介した映像や音響の特徴解析技術は，録画済みの番組以外にも，リアルタイム系への拡張により，スポーツ中継のハイライト生成への応用が考えられます．つまり，中継終了直後または数分程度の遅延でスポーツのハイライトを生成するわけです．通常サッカー中継においては，ハーフタイムにおいて前半のハイライトなどが放映されますが．この技術を適用すれば，ハーフタイムになると自動的に前半のハイライトが生成されているということも考えられます．

また，より番組の内容に踏み込んだ手法も検討が進められています．例えば，「話題分割技術」と呼ばれる手法があります[2]．ニュース番組では，一般に複数の出来事が報道されますが，それぞれのニュース項目が「話題」に相当します．ニュース番組からこのような話題の切替わり点を検出することにより，ニュース項目ごとに頭出しをすることができ，効率的な閲覧が可能になるほか，話題単位での検索など高度な映像アプリケーションが実現されます．

さらには，映像ばかりではなく，テキストやグラフィックスが縦横無尽に配置されているWeb画面を，携帯電話など小型な端末でも違和感なく，かつ操作性よく閲覧できるようにする技術[3]など，多種類の情報形式の特徴を活用し，時間・空間的にエッセンスを的確に抽出する手法が，日々考案されています．

参　考　文　献

1) M. Sugano, Y. Nakajima and H. Yanagihara: "MPEG Content Summarization Based on Compressed Domain Feature Analysis", IT Com 2003, SPIE 5242, 32, pp.280-288（Sep. 2003）
2) 帆足，菅野，内藤，松本，菅谷：“汎用的特徴量に基づく動画像話題分割手法”，信会論誌D，J89-D，10，pp.2305-2314（Oct. 2006）
3) 服部，松本，菅谷：“コンテンツ間距離の標準偏差に基づくWebページ動的分割方式”，情報処理学論誌，47，SIG 8，pp.81-89（June 2006）

滝嶋　康弘（たきしま　やすひろ）　1986年，東京大学電気工学科卒業．1988年，同大学院電子工学修士課程修了．同年，国際電信電話（株）（現（株）KDDI）に入社し，情報理論，画像符号化，低レートビデオ伝送方式などの研究，低レート符号化を応用したビデオ伝送システムをはじめとした応用システムの開発に従事．最近は，高度メディア解析技術に携わり，マルチメディアの解析・合成，自動理解の研究開発に従事．現在，（株）KDDI研究所知能メディアグループリーダ．工学博士．正会員．