オーディオビジュアル複合情報処理　35－5
(2001. 12. 13)

映像の特徴に応じたAVデータからの自動要約抽出方式に関する検討

菅野　勝　　中島康之　　柳原広昌

KDDI 研究所
〒356-8502 埼玉県上福岡市大原 2-1-15

E-mail: {sugano, nakajima, yanap}@kddilabs.jp

あらまし　MPEG 符号化されたオーディオビデオデータから，符号化データ上の特性を利用して要約情報を自動抽出する方式について検討した．先にオーディオ及びビデオの特性を併用して効率的な要約情報の抽出が可能になることを示したが，要約情報の判定に用いる閾値処理は，ビデオの動きなどの特徴が既知であることを前提としていた．本研究は，ビデオの特徴が未知の場合でも，その特徴に応じて閾値処理を適応的に変化させて要約情報を抽出する．また，要約情報の判定に用いるパラメータとしてコンテンツ記述の標準化 MPEG-7 で規定されている記述子を利用するため，より柔軟かつ MPEG-7 との親和性の高い方式を実現することが可能となった．

キーワード 要約情報，特徴記述，MPEG-7，記述子，類似ショット

A Study on Automatic Summary Extraction from AV Data According to Visual Features

Masaru SUGANO, Yasuyuki NAKAJIMA and Hiromasa Yanagihara

KDDI R&D Laboratories Inc.
2-1-15 Ohara, Kamifukuoka, Saitama 356-8502, Japan

E-mail: {sugano,nakajima,yanap}@kddilabs.jp

Abstract　We have already proposed the automatic summary extraction method from the MPEG coded audio-visual data, by using audio-visual characteristics on compressed domain. Our previous method assumes that the visual features such as motion are known a priori. In this paper, we show the extended method for extracting the audio-visual summary without any knowledge of the content by adaptively evaluating the visual features. Additionally, in order to extract summary, our method utilizes some descriptors defined in the MPEG-7 standard and can generate some MPEG-7 descriptions; our method is very flexible and can be easily applicable to the MPEG-7 applications.

Key words　Summary, Feature Description, MPEG-7, Descriptors, Similar Shot

1. はじめに

ISO/IEC で標準化が行われているコンテンツ記述標準の MPEG-7 や, PVR (Personal Video Recorder) による蓄積型テレビの標準を策定する TV-Anytime Forum などでは，マルチメディアコンテンツの概要を短時間で把握するためのメタデータ要素として，要約情報（サマリ，ダイジェスト）を記述するためのツールを定義している[1][2]．いずれも要約情報の抽出方法は標準規定外であるが，要約情報の記述にはその効率的な抽出が不可欠となる．

筆者らはこれまで，MPEG 符号化されたオーディオビデオデータから，オーディオやビデオの符号化データ領域での時空間的な特性を用いて，意味的に重要であると考えられる要約情報を効率的に自動抽出する方式について検討している[5]．この方式は，コンテンツの大まかな特徴（動的か静的かなど）が既知である場合に，符号化データから自動且つ低コストに要約情報を抽出できる．また，外部から指定された長さの要約情報をコンテンツ全体から均一的に抽出することができるため，利用者の好みの長さでコンテンツ全体の概要把握を行うことができる．

本稿では，より多くの重要な特性を適用して抽出精度向上を図ると同時に，MPEG-7 への応用を考慮した方式について検討したので報告する．具体的には，要約情報の抽出処理の過程で MPEG-7 の一部の記述子を抽出し，これらを要約情報抽出の判定に用いることによって効率的な抽出を行える．また，それらの記述子を要約情報記述と共に MPEG-7 ファイルとして出力することが可能なため，効果的なコンテンツ記述が可能となる．更に，MPEG-7 記述子を判定要素に用いていることから，該当する記述子を含む MPEG-7 記述ファイルが存在すれば，そのファイルを用いて要約情報の抽出及び記述も可能である．

2. 提案方式の概要

まず，先に提案した方式[5]の処理手順を説明する．本方式では，以下の手順により要約抽出を行う．図 1 に処理の概要を示す．

1. コンテンツをショット検出によりショットへ分割する．この際，人間の脳で完全に処理できるショットの最小長（=3.5 秒[6]）以下の長さを持つショットは，要約情報の候補から除外する．
2. コンテンツ長を *CL*, 外部から指定される要約情報長を *SL*, 及び全ショット数を *NS* とし，分割数 $NP=NS\times SL/CL$ でコンテンツを等分割する．
3. 等分割した区間に属するショットに対し，オーディオ特性及びビデオ特性を解析する．
4. 解析結果に基づいて，等分割区間内での要約情報となるショットを決定する．
5. 要約情報と判定されたショットを集約する．

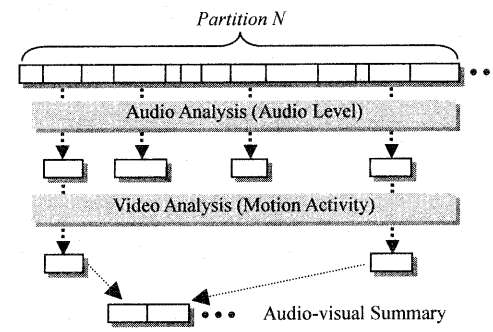

図 1　既存提案方式における要約情報構成

3. 提案方式の拡張

3.1 ビデオ特性の利用

先に提案した方式では，ビデオ特性について「動きアクティビティ」を定義し，判定要素として用いた．動きアクティビティは，MPEG ビデオデータに含まれる動きベクトルの大きさを用いて計算される．

本稿ではその拡張として，(1) 動き情報をより詳細に解析することと，(2) 反復ショット（類似ショット）を検出することを主なタスクとする．(1) については動きベクトルの大きさだけでなく方向や時間的な分布を考慮に入れることによって，より詳細なビデオ特徴の分類が可能となる．(2) については，類似ショットが複数存在する場合に，ビデオ特性を用いて類似ショットを検出する．これにより，ニュースなど類似ショット（アンカーショット）が重要なコンテンツではそれらのショットを集約でき，スポーツなど類似ショットが多数存在するコンテンツでは類似ショットを除外することによって特徴的なショットのみを抽出することができる．

両者を効果的に達成すること，及びコンテンツ記述への応用も視野に入れ，MPEG-7 Visual[3]で定義されている動き記述子を抽出すると共に，類似検索の精度を向上させるため色特性も併用し，同様に MPEG-7 で定義されている色記述子を抽出する．ここでは，それぞれ符号化データ領域の簡単な演算で抽出及び類似度の判定が可能であり，映像の類似検索としての精度が高い「動きアクティビティ記述子

(Motion Activity Descriptor)」と「色配置記述子(Color Layout Descriptor)」を利用する.

3.1.1 動きアクティビティ記述子

動きアクティビティ記述子を用いることにより,動きの特性(強度,方向,時空間的な分布など)を把握してより詳細なビデオ特徴の分類を行うと同時に,類似検索によって類似ショット検出を行う.MPEG-7で規定されている動きアクティビティはいくつかの要素から構成され,Pピクチャにおけるマクロブロックの動きベクトル値に基づいてフレーム毎に計算される.ショットが持つ動きアクティビティは,ショット内の代表フレームにおける動きアクティビティを用いることもできるが[4],ここではショット内に含まれる全てのPピクチャの平均値を採用する.本方式では必須要素である動き強度でビデオ特徴を解析し,類似度の判定に最頻方向を用いる.

- ◆ 動き強度
  フレーム内の全動きベクトルの大きさに関する標準偏差を求め,5段階に量子化した値(1:Very Low Activity〜5:Very High Activity).
- ◆ 最頻方向
  フレーム内の全動きベクトルの角度成分を求め,8方向(45度刻み)のうち最頻の方向を示す値.

3.1.2 色配置記述子

類似ショット検出を行うため,上記の動きの特性のほかに,色の特性を用いる.色配置記述子は,周波数領域で色情報の空間的な配置を表すものである.ショットに対する色配置記述子の値は,ショット内のキーフレームなどの代表フレームにおける値で表すことができるが,本方式で採用しているショット検出アルゴリズム[7]においては,簡易動き補償を用いたDC画像を作成する処理が含まれているため,カット点と判定されたDC画像(即ちショットの先頭画面)から,色配置記述子を抽出する.

図2に示すように,オリジナルのサイズに対して水平垂直方向に1/8されたDC画像を8×8のブロックに分割する.画像の水平/垂直サイズが8の倍数でない場合は,右側及び下側のブロックは他のブロックよりも画素数を減らし,補間は行わない.次に,この8×8ブロックのそれぞれにおいて,平均値を求め占有色(Dominant Color)とみなす.更にこのブロックの輝度及び色差成分に対して8×8のDCTを施し,低域側からある帯域までのDCT係数を取り出してDC成分,AC成分をそれぞれ非線形量子化する.

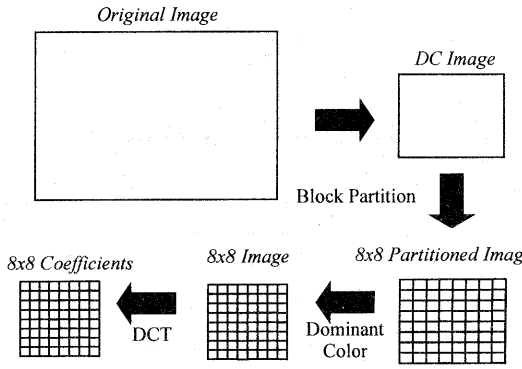

図2 符号化データからの色配置記述子の抽出

3.2 オーディオ特性の利用

先に提案した方式は,オーディオの特性としてMPEG符号化データ領域でのサブバンドエネルギーをサブバンドごとに重み付けし,得られた値をオーディオレベルとみなし,レベルがある一定以上の場合に要約情報の候補としていた.先の実験により,オーディオレベルは要約情報を抽出する際に,意味的な面で大きく寄与することが分かっている.しかし,複数のショットに渡って音声が連続しているシーンが含まれていることがあり,このような場合にはショット単位での要約情報が必ずしも適しているとは言えない.

そこで拡張方式では,より意味的な側面を重視するため,「音声」によるオーディオカット点を優先する処理を行う.これは,図3のようにビデオのショット区間とオーディオの音声区間が同期していない場合,音声クラスが他のクラスへ変化する点をショット終了点と見なし,要約情報の候補として抽出する区間の終了点を拡張する.これにより,音声が途切れることなく,より自然な要約情報の抽出を可能とする.但し,ビデオ特性としての動きアクティビティや色配置などの特徴値の更新は行わない.

また,コンテンツの特徴により優先するクラスを適応的に変化させる.ここでのクラスとは,文献[8]に示した方式により分類されるクラスのうち,「音声」及び「音楽」である.具体的には,ショット毎のクラス分布からそのショットにおける優位クラスを求め,これに基づいて等分割区間内での優位クラ

スを決定する．あるショットのクラスが等分割区間内の優位クラスに一致しない場合，要約情報の候補から除外する．尚，「歓声雑音」については，特にスポーツなどにおいて重要な要約情報となり得るため，他クラスと比較してより優位なクラスであるとする．

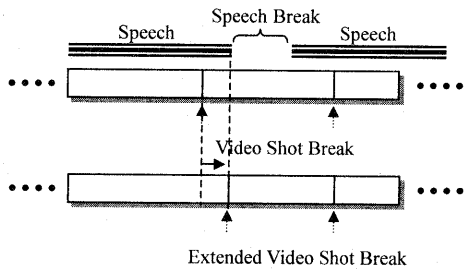

図 3　オーディオとビデオの非同期性

3.3 ビデオ特徴に応じた適応閾値処理

先に提案した方式では，コンテンツのビデオ特徴（内容）が既知であることを前提とし，各等分割区間内でビデオ特性（動きアクティビティ）をどのように閾値処理するかは予備知識により決定していた．例えば，ドキュメンタリーなどの動きが小さいショットが重要な場合は，等分割区間内で決定した閾値よりも小さい値を持つショットを要約情報の候補とし，スポーツなどの動きが大きいショットが重要な場合は，等分割区間内での閾値よりも大きい値を持つショットを要約情報の候補とすることにより，適切な要約情報の抽出が可能なことを示した．更に，区間毎に閾値を適応的に変化させることで，ビデオ特徴の異なるクリップが混在していても，効果的に要約情報を抽出できることが分かった．しかし，例えば長尺のコンテンツなどではその内容が未知の場合にも効果的に要約情報が抽出できることが望ましい．従って，ビデオの特徴値を用いて自動的にビデオ特徴を解析し，最適な閾値処理を行う必要がある．

そこで，3.1.1で述べた動きアクティビティの動き強度を用いて，ビデオ特徴の解析を行う．ショット内のPピクチャでの動き強度 I_F を求めた後，ショット全体における動き強度の平均 I_S を求め，更に分割区間内での動き強度の平均 I_P を求める．前述したように，フレームにおける動き強度 I_F は 1〜5 のいずれかのレベルとして分類され，レベルが大きいほど動き強度が大きいことを示す．ここでは I_S の値が 2.5 である点（Low〜Medium の中間）を境界として，動きが大きいものと小さいものに分類する（図 4）．

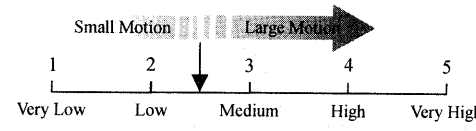

図 4　動き強度によるビデオ特徴の分類

分割区間内の動き強度の平均 I_P が 2.5 より大きい場合，その分割区間内に含まれるショットは動的なショットであると見なし，要約情報の候補となっているショットの平均動き強度 I_S が分割区間内の平均動き強度 I_P 以上の場合に，該当するショットを要約情報の候補とする．

逆に，分割区間内の動き強度の平均 I_P が 2.5 未満の場合，その分割区間内に含まれるショットは静的なショットであると見なし，要約情報の候補となっているショットの平均動き強度 I_S が分割区間内の平均動き強度 I_P 未満の場合に，該当するショットを要約情報の候補とする．

また，平均ショット長などもビデオ特徴の解析に利用することができると考えられる．

以上を考慮した要約情報抽出の拡張方式の処理フローを図 5に示す．

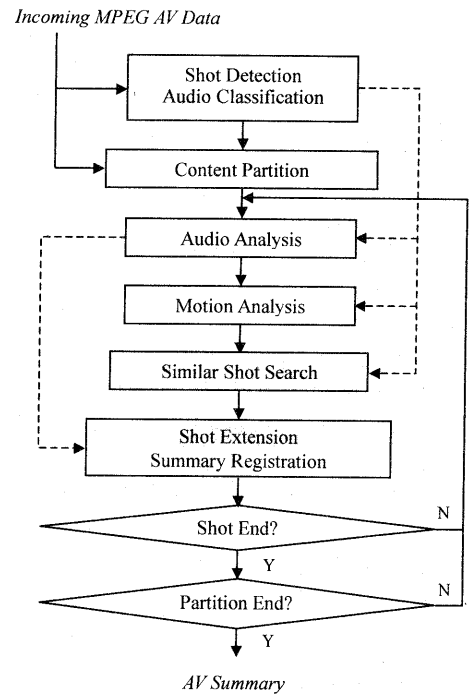

図 5　拡張方式の処理フロー

4. 検証実験

4.1 実験の手順

MPEG-1 符号化されたオーディオビデオコンテンツ（SIF）を用いて検証実験を行った．使用したコンテンツは①テレビのキャプチャ映像 60 分と文献[5]でも用いた②テレビ番組 22 分であり，①はスポーツ，ニュース，情報番組，②はドキュメンタリーとスポーツというように，一つのコンテンツ内にビデオ特徴の異なる複数のクリップが含まれている．このコンテンツから，それぞれ 1/10，1/5 の長さを持つ要約情報を抽出する実験を行った．抽出された要約情報としてのショットの開始時間と継続時間を記録し，ASX（Active Streaming XML）形式のファイルを作成して Windows Media Player で再生，確認を行う．

①のコンテンツについては，自動ショット検出により 320 のショットを検出し，うち 3.5 秒以下のショット（28 ショット）は予め候補から除外しておく．コンテンツの分割数は 1/10 要約長の場合で 320×360/3600=32 となる．これらの区間から，平均ショット長 3600/320≒11 秒程度のショットを抽出する．

まずオーディオ特性を解析し，オーディオレベルが分割区間内で相対的に大きく，且つ優位オーディオクラスに一致するものを要約情報の候補とみなす．次にビデオ特性については，動き強度を判定要素として用い，3.3に示したように分割区間内の平均動き強度 I_P と各ショットの動き強度 I_S を比較して要約情報の候補を絞り込む．また3.2に示したように，候補となった要約情報が音声クラスを持つ場合，音声区間終了点と同期してショット終了点を拡張する．

更に，3.1に述べた動き特性及び色特性の側面から類似検索を行うために，要約情報の候補としてのショットが決定された場合，前者について動き強度・最頻方向を，後者について色配置を用いて以降の全ての分割区間内に含まれる要約情報候補のショットと比較を行い，類似ショットを検出する．

4.2 実験結果と考察

図 6に，コンテンツ①から抽出したショットの平均動き強度 I_S，分割区間内の平均動き強度 I_P を示す．破線は 1/10 要約長での分割区間である．実験は予備知識がない状態で行ったが，コンテンツ①は 33 分付近（約 2000 秒）までがスポーツ（テニス），その後ニュースが 5 分挿入され，残りが情報番組である．

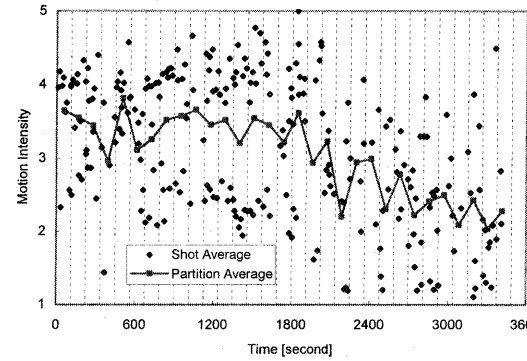

図 6　コンテンツ①の動き強度分布

図 6から分かるように，スポーツ部分では分割区間内の平均動き強度 I_P が比較的大きい値を示し，ショット内の平均動き強度 I_S では動きの大きい箇所と小さい箇所が明確に分かれている．これはテニスというスポーツの性質上，「静→動→静」が比較的短い間隔で繰り返されるためで，ビデオの特徴が適切に反映されている．これに対し後半のニュース，情報番組の部分では分割区間内の平均動き強度 I_P は全体的に低下し，ショット内の平均動き強度 I_S は様々な範囲に分散しており，ショット長も比較的長い．

このようなビデオ特性を評価すると，動き強度の閾値を 2.5 とした場合，情報番組の一部の区間が静的であると見なされ，その他全ての区間は動的であった．これに応じて閾値処理を適応的に変化させたところ，要約情報として適切なショットが抽出された．文献[5]の方式ではビデオ特徴が予め分かっている状態で閾値処理を決定していたが，本方式ではコンテンツに関する予備知識がなくても，ビデオ特徴に応じた要約情報の抽出が可能となった．

類似ショット検出については，特に色特性による検出が有効であり，本実験では類似ショットを除外した．これによりスポーツの反復ショットや情報番組のスタジオショットを除外できたが，同時にニュースのアンカーショットも除外されてしまうため，検出した類似ショットを除外するか採用するかについては，更に詳細なコンテンツ解析が必要である．

またオーディオ特性を用いた結果，前半では歓声雑音を含むショットが抽出された．スポーツでは試合などで何らかの変動の直後に歓声雑音が起こるため，オーディオ特性が有効である．このとき，実際のハイライトは歓声雑音の直前に位置するため，ビデオ特徴の評価によりスポーツと分類される場合，歓声雑音の直前のショットを抽出するなどの適応的

な処理が必要であると考えられる．また，スポーツ解説やニュース，情報番組においては，ビデオと音声区間の非同期が多く存在したため，音声区間によるショット拡張が効果的に作用し，音声の途切れなくより自然な要約情報を構成できた．但し，指定された要約長を大きく超える場合があるため，予め音声区間を考慮に入れておく必要がある．尚，1/5 要約は 1/10 要約を含む形で補足的なショットが得られた．

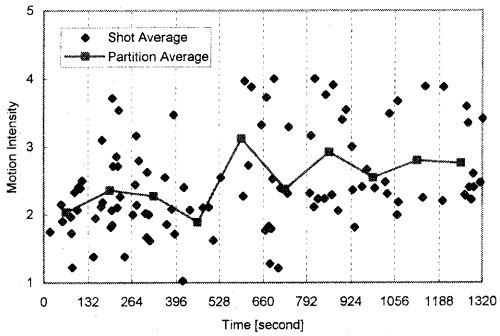

図 7　コンテンツ②の動き強度分布

図 7はコンテンツ②から抽出したショットの平均動き強度 I_S及び分割区間内の平均動き強度 I_Pである．②では後半 15 分がスポーツ（サッカー）であるが，スポーツとしての動き強度の特性は①と類似している．また，分割区間内の平均動き強度 I_Pもクリップの特性を反映していることが分かるが，①と比較して全体的に値が低い．これらの値はエンコーダに依存すると考えられるため，動き強度の閾値処理についてはコンテンツごとに対しても適応的に変化させることが有効である．コンテンツ②の後半は歓声雑音を用いるだけでハイライトの構成が可能であった．

5. MPEG-7 記述への応用

本稿で述べた要約情報抽出方式は，その過程において MPEG-7 で定義されている一部のビデオ特徴値を抽出し，判定に用いていることから，入力オーディオビデオコンテンツから一部の MPEG-7 記述を生成するのに有効である．具体的にはショット検出，動き／色特徴値抽出，及び要約情報の決定であり，これらに対応した記述を生成することができる（図 8）．本方式では，これらの特徴値は全て符号化データ上で抽出しており，抽出コストも抑制できる．

逆に，本方式による要約情報抽出で必要な特徴値が含まれる MPEG-7 記述があれば，その記述データを基に要約情報の抽出及び記述を行うことができる．

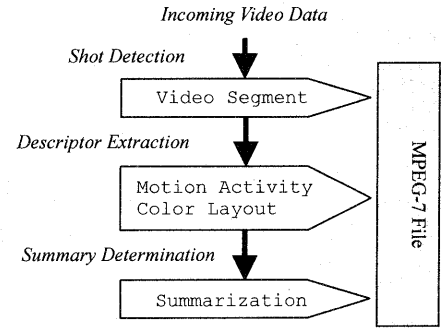

図 8　本方式の MPEG-7 記述への適用

6. まとめ

本稿では，MPEG 符号化されたオーディオビデオデータから，オーディオやビデオの符号化データ上での特性を用いて要約情報を抽出する方式について，コンテンツのビデオ特徴が未知の場合でも，特徴パラメータの値や変動などを考慮することによって効果的に要約情報を抽出することが可能であることを示した．また，コンテンツ記述標準である MPEG-7 との連携を考慮し，MPEG-7 で定義されている記述子を要約情報の抽出に用いることによって，MPEG-7 との親和性の高い方式を実現できる．

謝　辞

日頃ご指導頂く KDDI 研究所の浅見所長，並びに松島副所長，松本執行役員に感謝致します．

文　献

[1] ISO/IEC FDIS 15938-5, Information Technology – Multimedia Content Description Interface – Part 5: Multimedia Description Schemes, October 2001

[2] TV-Anytime Forum, “Specification Series: S-3 on Metadata”, SP003V11, August 2001

[3] ISO/IEC FDIS 15938-3, Information Technology – Multimedia Content Description Interface – Part 3: Visual, July 2001

[4] ISO/IEC JTC1/SC29/WG11/N4362, “MPEG-7 Visual part of eXperimentation Model Version 11.0”, July 2001

[5] 菅野，中島，柳原，「オーディオビデオの特性を用いた自動要約抽出方式に関する検討」，情処研報，AVM-32-1，pp.1-6，2001

[6] J. Saarela and B Merialdo, “Using content models to build audio-video summaries”, SPIE Conf. On Storage and Retrieval for Image and Video Database VII, 1999

[7] 中島，氏原，米山，「部分復号を用いた MPEG データからのカット点検出」，信学論 D-II，Vol.J81-D-II，No.7，pp.1361-1371，1998

[8] 中島，陸，菅野，柳原，米山，「MPEG 符号化データからのオーディオインデキシング」，信学論 D-II，Vol.J83-D-II，No.5，pp.1564-1575，2000