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

菅野　勝　　中島　康之　　柳原　広昌

KDD 研究所

〒356-8502　埼玉県上福岡市大原 2-1-15

sugano@kddlabs.net

ディジタル放送の開始などにより，MPEG 等で圧縮されたコンテンツの普及が進み，HDD を利用した PVR（Personal Video Recorder）も登場している．このようなディジタルコンテンツを今後利活用するためには，その内容を短時間で把握するための要約情報が重要となる．本稿では，MPEG で圧縮されたオーディオビデオコンテンツの符号化データ上の特性を解析し，自動且つ低コストな処理により要約情報を抽出する方式について検討したので報告する．提案方式を適用した検証実験において，オーディオとビデオの特性を効果的に利用することによって，コンテンツの概要を把握できる要約情報を抽出できることを確認した．

A Study on Efficient Summary Extraction using Audio-visual Features

Masaru Sugano　Yasuyuki Nakajima　Hiromasa Yanagihara

KDD R&D Laboratories Inc.

2-1-15 Ohara, Kamifukuoka, Saitama 356-8502, JAPAN

sugano@kddlabs.net

Vast of digital audio-video contents such as MPEG compressed audio-visual data can be found in, for example, a digital broadcasting, and PVR (Personal Video Recorder) has been developed using HDD. It is very useful to create summary information in order to roughly grasp the content of the audio-visual data, in case of handling or re-using these contents. In this paper, we propose an automatic summary extraction method that efficiently employs certain properties of compressed audio-visual data. The proposed algorithm analyzes audio-visual features on compressed domain, and enables automatic and low-cost extraction of audio-visual summary data. Verification experiment has shown that the summary representing the rough content can be extracted.

1. はじめに

ディジタル放送の開始などにより，MPEG 等で圧縮されたオーディオビデオコンテンツの普及が進んでおり，HDD を利用した PVR (Personal Video Recorder) も登場しつつある．また，DVD などのパッケージによる MPEG ビデオも普及しつつあり，広帯域インターネットでの映像配信なども注目を集めている．今後，このようなディジタルコンテンツの利活用には，ビデオやオーディオの内容を短時間で把握するための要約情報（サマリ）が重要になると考えられる．

一方，マルチメディアコンテンツの効率的な検索などを目的としたコンテンツ記述方式として，ISO/IEC において MPEG-7 標準化が行われており，要約情報の記述スキーム (Description Schemes) が標準に規定されている[1]．MPEG-7 では記述する特徴値の抽出方法は標準外としているため，要約情報の抽出方法は規定されていないが，コンテンツの閲覧などに有効な要約情報の記述には，その効率的な抽出が不可欠となる．

本稿では，MPEG で圧縮されたオーディオビデオコンテンツの符号化データ上の特性を解析し，自動的且つ低コストに要約情報を抽出する方式[2]について検討したので報告する．

2. 既存の自動要約抽出技術

これまでに，自動要約技術に関しては幾つかの報告がある．例えば，ビデオのシーン変化点を検出した後，階層構造化を行い，各シーンに優先度を付与することによってビデオの要約を自動的に作成する方法が報告されている[3]．階層構造において上位階層ほど優先度が高くなるように設定されるが，階層構造化は手動で行う必要があり，その処理に多くの時間を要する可能性がある．また，優先度の付与はシーンが属する階層に依存して行われるため，実質的には人手を要することが多い．

また，汎用的なビデオの要約情報の作成を，制約つきの最適化問題として捉えている報告もある[4]．制約としては，最小のショット長，オーディオとビデオの同期，ビデオの連続性，及びオーディオとビデオの冗長性などである．そして，手動によってビデオの内容モデル（記号的な記述）を構築し，要約情報の作成を行っている．ビデオの内容モデルの構築を手動で行わなければならないほか，ビデオのセグメントの分類においては，圧縮データ上では実現が困難である高度な認識技術などが必要となるため，要約情報抽出に要する処理が大きくなることが予想される．

文献[5]では映画の予告編に特化した要約情報の作成を目的としている．主な作成手順としては，①ビデオのショットへの分割，②特別なイベントを含むクリップの解析，③クリップの選択，及び④クリップの集約である．ショットから抽出される特別なイベントとしては，俳優の顔の認識・会話の識別，タイトルやテロップからの文字情報の抽出，及び銃撃や爆発などであるが，この方式でも上記の[4]と同様にして，コンテンツの意味内容にまで立ち入った高度な処理が必要となる．

このように，従来技術ではオーディオビデオコンテンツの入力から要約情報を抽出するまでの過程において，手動による処理や高コストな処理が必要となることが多い．また，場合によっては意味的に重要な内容を含むオーディオ情報を効果的に利用しないため，適切な要約情報が得られないことも考えられる．また，要約情報の構造的な側面から見ても，オーディオビデオコンテンツ全体から均一に要約情報が抽出される保証はないため，要約情報として必ずしも適さない場合があると考えられる．

3. 提案方式

本稿では，2. で述べたような手動による処理や高コストな処理を必要とせず，自動且つ低コストで圧縮オーディオビデオコンテンツから要約情報を抽出する方式を提案する．これらは，MPEG で圧縮されたオーディオビデオコンテンツにおいて，圧縮データ上でオーディオ及びビデオの時空間的な特性を解析し，それらを統合的に評価することによって，要約情報の候補としてのオーディオビデオ区間を決定するものである．また本方式は，コンテンツ全体から均一に要約情報を抽出できるほか，外部から与えられる要約情報長に近い長さを持つ要約情報を構成することが可能であり，例えば 60 分のコンテンツを 15 分に要約したり，5 分に要約したりすることができる．

本方式では，具体的な処理としてまず①オーディオビデオコンテンツをショットへ分割し，②コンテンツをある基準により等分割する．そして，等分割した区間に属する各ショットに対して，③オーディオデータの解析，及び④動き情報の解析を行い，これらの解析の結果要約情報として見なされたショットを⑤要約情報として決定，集約する．以下で，これら処理についてより詳細に説明する．図 1は本方式の処理フローを表している．

3.1 ショット分割

本方式は，ショット単位で要約情報の候補か否かを決定する．よって，第一の処理として，コンテンツ全体をショットへ分割する．ショット検出は，文献[6]に示されている DC 画像の差分と色差相関を利用するアルゴリズムを用いて行う．またこのとき，ショット数 NS をカウントする．

尚，人間の脳で完全に処理できるショットの最小の長さは 3.5 秒であるとの報告があるため[4]，これ以下の長さを持つショットについては，要約情報の候補から除外する．

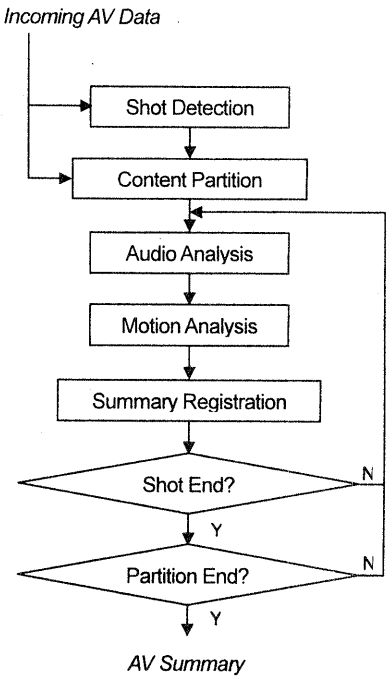


図 1　提案方式の処理フロー

3.2　コンテンツの等分割

コンテンツ長 *CL*, 外部から指定される要約情報長 *SL* 及び3.1で得られたショット数 *NS* から分割数 *NP* (=*NS*×*SL*/*CL*) を求めてコンテンツを等分割する．これにより，コンテンツ全体から均一に要約情報を抽出でき，ある特性を利用したときに，特定の部分から要約情報が集中的に抽出されることを回避する．但し，*SL*>*CL*/*NS* であるとする．

また，例えば文献[5]では，映画のクライマックスシーンとしてコンテンツ最後尾の 20%の区間を意図的に除外しているが，本方式のように区間分割を行うことによって，要約情報を抽出する区間を適応的に選択することができる．

3.3　オーディオの解析

本方式ではオーディオの特性を積極的に利用する．これは，例えばドキュメンタリーなどにおけるナレーション及び付随する BGM や，スポーツなどにおける歓声，映画などにおける銃撃や爆発などのイベントが，意味的に重要なことが多く，要約情報の構成に有効であると考えられるためである．また，オーディオ単体で考えた場合でも，例えば楽曲の主題部分（サビ）は比較的レベルが大きいことが多く，レベルの大きいオーディオ区間を聴くことにより楽曲の曲調など，ある程度の内容を把握することができると考えられる．

本方式では，MPEG で符号化されたオーディオのサブバンドデータを判定要素として利用し，上記等分割区間に属するショットが無音または十分小さいレベルを持つ場合にこれらのショットを要約情報から除外する．また，ある一定のレベル以上のレベルを有するオーディオ区間を持つショットを要約情報の候補とする．従って，上記のような意味的に重要なショットが抽出の対象となる．

無音区間の判定方法については，筆者らは既にサブバンドエネルギーの分散を用いた方式を提案している[6]．これに対して本方式では，無音だけでなくオーディオのレベルも評価するため，以下のような手順により MPEG オーディオのサブバンドデータを用いてオーディオレベルの判定を行う．

① ショット *S* のオーディオ部分から，各オーディオフレームにおけるサブバンドデータを抽出し，サブバンドエネルギー *sb* を求める．
② サブバンドにより重み付けされた，あるフレームにおけるサブバンドエネルギーの総和 *SBE* を計算する．
③ *SBE* に基づいて単位時間当たりに含まれるオーディオフレームでのサブバンドエネルギー総和 SB_S を計算する．
④ 各ショットにおける平均サブバンドエネルギー総和 ASB_S を求める．これを，オーディオレベルとして評価する．

1 秒当りの単位時間サブバンドエネルギー SB_S 及び各ショットでの平均サブバンドエネルギー総和 ASB_S は，以下の式（2）ようにして求める．

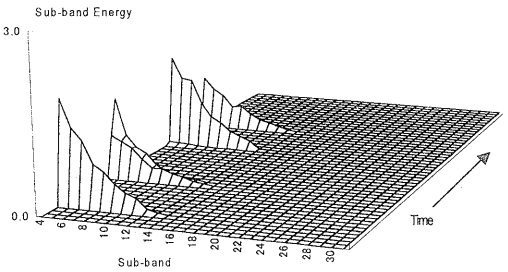
$$SBE = \sum_{k=0}^{31} \alpha_k \times sb_k$$
$$SB_S = \sum_{k=1}^{NAF} SBE \qquad (2)$$
$$ASB_S = \frac{SB_S}{SHL}$$

ここで，α_k はサブバンド k に対する重み付け，*NAF* は 1 秒当りのオーディオフレーム数，sb_k はフレームにおけるサブバンド k のエネルギー，*SHL* は対象となるショットの長さ（秒）である．

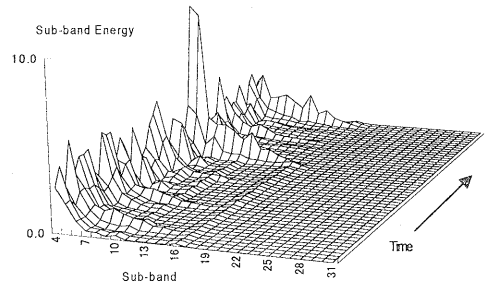
α_k ついては，無音，音声，音楽，効果音などのカテゴリを総合的に判定するため，予備実験の結果により次のように定めた．MPEG オーディオから得られる 32 のサブバンドデータのうち，第 0 サブバンドは全てのカテゴリに共通の基本バンドであるとみなし，全体に対して 0.1 の重み付けを与える．また，第 1～第 3 サブバンドについては，主に音声（声楽を含む）が寄与すると考えられるため，それらの和に対して 0.2 の重み付けを行う．そして，主に効果音（図 2(a)）や楽曲（図 2(b)）などの影響を受ける残りのサブバンド（第 4～第 31）の総和に対して 0.7 の重み付けを行う．従って，*SBE* は次の式（3）で与えられる．

$$SBE = 0.1 \times sb_0 + 0.2 \times \sum_{k=1}^{3} sb_k + 0.7 \times \sum_{k=4}^{31} sb_k \quad (3)$$

このようにして求められた各ショットにおけるオーディオレベルが，予め定められた閾値よりも小さい場合は，無音または十分小さいレベルとみなして要約情報の候補から除外する．逆に，各分割区間内での最大値から順に，相対的に大きな値を持つショットを，要約情報の候補とみなす．



(a) 効果音（マシンガン音）



(b) 楽曲（ロック音楽）

図 2　サブバンド4〜31のエネルギー分布

3.4　動き情報の解析

本方式では，ビデオにおけるカメラやオブジェクトの動きなどに着目し，各ショットにおける動きの度合いを示す動きアクティビティを定義して，要約情報の判定に用いる．例えば，スポーツなどでは動きが大きいショットが重要な意味を持つことが多く，ドキュメンタリーなどではカメラが固定されたショットやカメラの動きが静止したショットが重要な意味を持つことが多いと考えられる．

動きアクティビティは，ビデオデータに含まれる動きベクトルを用いて計算される．各ショットに対する動きアクティビティは，具体的には以下の手順により求める．式（4）に計算式を示す．

① ショット S に属する全ての予測符号化フレームから，ある大きさよりも大きい動きベクトルを持つマクロブロックをカウントし，該当する動きベクトル情報 MV を抽出する．

② 動きベクトル情報 MV を用いてショット S 全体の動きアクティビティ総和 MA_S を計算する．

③ ショット内動きアクティビティ総和 MA_S をショット内の予測符号化フレーム数 NPF_S で割ることにより，ショット内平均動きアクティビティ MA を計算する．

$$ASMV = \sum_{n=1}^{NMB_X} |MV| \quad |MV| > X$$
$$MA_S = \sum_{m=1}^{NPF_S} \frac{ASMV}{NMB_X} \quad (4)$$
$$MA = \frac{MA_S}{NPF_S}$$

ここで，$ASMV$ は大きさが X より大きい動きベクトルのフレーム内絶対値総和，NMB_X はフレーム内での大きさが X 以上の動きベクトルを持つマクロブロック数である．本方式では X=1 を適用し，順方向動きベクトルのみを用いた．

上記で求められた動きアクティビティを，各分割区間内で決定する閾値により閾値処理する．このとき，前述のようにスポーツなどでは大きい動きアクティビティを，ドキュメンタリーなどでは小さい動きアクティビティを要約情報として採用するなど，コンテンツにより適応的に変化させる．

3.5　要約情報の決定及び登録

各等分割区間から，上記の条件に適合するショットを決定する．決定されたショットが平均ショット長（$=CL/NS$）以上であれば処理を終了して該当する区間の要約情報として登録し，次の区間の処理へ移行する．一方，決定されたショットが平均ショット長に満たない場合は，次の候補を採用し，合計のショット長が平均ショット長を超えるまで繰り返す．このようにして各区間から抽出された全てのショットを結合し，要約情報を構成する．図 3に，本方式による要約情報の構成を示す．

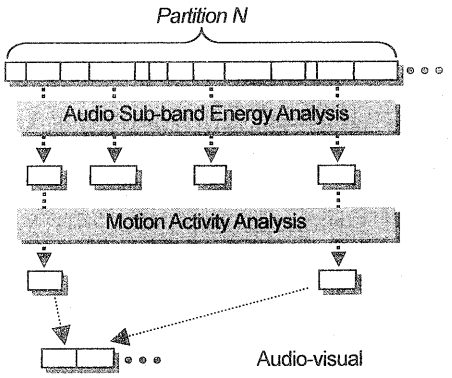


図 3　要約情報の構成

4. 検証実験

4.1 実験の手順

本方式の有効性を検証するために，MPEG-1 のSIF レベルで符号化されたオーディオビデオコンテンツ（約 9 分（=544 秒）のドキュメンタリー映像）について，元コンテンツに対して 1/10 及び 1/5 の長さを持つ要約情報（以下，1/10 要約長，1/5 要約長）を抽出する実験を行った．

本方式により要約情報として決定されたショットについて，各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 により要約情報の再生，確認を行う．

このコンテンツは目視により全体で 46 個のショットが認められるが，3.1に示した自動ショット検出により 40 個のショットを検出し，うち 3.5 秒以下のショット（1 ショット）は予め候補から除外する．尚，過剰検出は 2 ショットであった．分割数 NP は，1/10 要約長の場合，

$$NP_{1/10} = 40[\text{ショット}] \times 54.4[\text{秒}] / 544[\text{秒}] = 4$$

と決定され，1/5 要約長の場合は $NP_{1/5} = 2 \times NP_{1/10} = 8$ と決定される．これらの区間から，平均ショット長，即ち 544/40≒14[秒]程度のショットを抽出する．

要約情報の抽出については，その定量的な評価方法が確立されていないため，参照方式による抽出結果との主観的な比較により方式の評価を行った．参照方式として，図 4に示すように各等分割区間の開始点から一定の長さを抽出して要約情報を構成した．1/10 要約長の場合，分割区間数が 4，要約情報長が 54.4 秒であることから，各分割区間の開始点から平均ショット長（=14 秒）と同じ長さの区間を抽出した．同様にして，1/5 要約長に対しては 8 区間の開始点から抽出した．但し，ショットの開始，終了とは同期させていない．

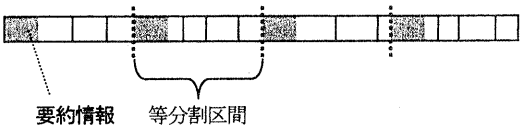


図 4 要約情報抽出の参照方式（NP=4）

4.2 実験結果と考察

図 5は，1/10 要約長の場合の，各ショットにおける平均オーディオレベルと平均動きアクティビティをプロットしたものである．x 軸は時間[秒]を表し，グラフ左側の y 軸はオーディオレベルの値，グラフ右側の y 軸が動きアクティビティの値を表す．尚，グラフにおける破線は 1/10 要約長における分割区間の区切りである．本方式により抽出されたショットを実線矢印で示す．

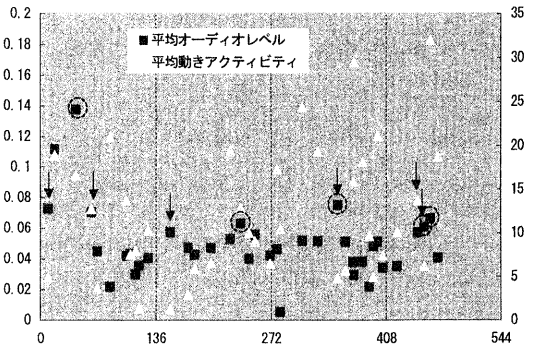


図 5 オーディオレベルと動きアクティビティ

平均オーディオレベルに関しては，区間ごとに分散が異なっている．特に，第 1 区間の前半部分は他のどの部分よりも大きい．これは，オープニング部分でBGMとナレーションが入っていることによる．動きアクティビティについては，第 2 区間で静止したショットが多くなっている．

本方式により抽出した要約情報と図 4の参照方式により抽出した要約情報を比較した結果，上記3.1のショット分割，3.2のコンテンツ等分割，及び3.3のオーディオレベル解析を適用するだけでも（図中破線円），十分コンテンツの概要が理解できるレベルの要約情報が抽出でき，オーディオが意味的な重要性に大きく寄与することが分かった．但し，3.2を適用しない場合には，図 5からも分かるようにオーディオレベルの大きい値が第 1 区間に集中しているため，この区間から要約情報が抽出されてしまい，コンテンツの等分割も有効な手段であることが分かった．更に，3.3と3.4を併用することによって，より適切な要約情報の構成が可能になることを主観評価により確認した．尚，実験に用いた映像はドキュメンタリーなので，分割区間内の平均動きアクティビティよりも小さい動きアクティビティを持つショットを要約情報として抽出した．

一方，1/5 要約長については，1/10 要約長で抽出されたショット含んでいるだけでなく，1/10 要約長の意味内容を補足するような比較的重要なショットが追加されているため，より詳細な内容を把握することが可能となった．

尚，3.5に示した方法により各分割区間でのショット長を決定したところ，最終的に得られた要約情報長は，1/10 要約長で 68 秒，1/5 要約長で 128 秒であった．ここで用いたコンテンツは比較的ショット長の分散が少ないが，要約情報の候補となったショットが平均ショット長と比べて極端に長い場合には，所望の要約情報長に近づけるため適宜ショットの中断なども必要となる．

図 6は，別の MPEG-1 コンテンツに対して，図 5と同じデータをプロットしたものである．これは 22 分のコンテンツであり，前半 7 分のインタビューと後半 15 分のスポーツ（サッカー）から構成

されている．グラフ中の2点鎖線は2つのシーンの区切りを表し，分割点は 1/10 要約長に対応する．抽出されたショットを実線矢印で示す．この結果，サッカーでのゴールシーンも含めコンテンツの概要が十分把握できる要約情報が抽出された．このように，同一コンテンツ内の異なるシーン間で動きアクティビティの分布が大きく変化する場合，動きアクティビティの大小いずれを優先するかを適切に決定することが有効であるが，更に同一シーン内でも区間によって動きの分布が異なることがあるため，シーン内で単一の閾値を用いるよりも，各区間内での動きアクティビティの分布に応じて，区間毎に閾値を動的に変化させることで，より適切な要約情報が抽出できる．

また，第4・第6区間でオーディオレベルが突出しているが（図中破線円），いずれもサッカーのゴールシーンに相当し，オーディオ情報を用いるだけでハイライトを抽出することができる．これらは同時に動きアクティビティも大きく，その他シュートやフリーキックなどの重要なショットについても，大きい動きアクティビティを採用することによって抽出することが可能となる．

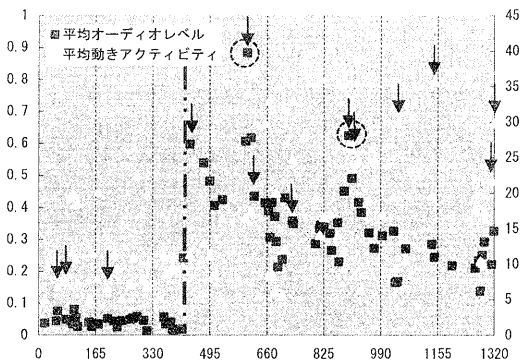


図 6　オーディオレベルと動きアクティビティ
（インタビュー＋スポーツ）

4.3　今後の課題

本実験では，ショット単位で要約情報を抽出しているが，図 7に示すように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．これについては，例えばショットが切り替わってもオーディオが連続していれば，オーディオのカット点を優先するなどの考慮が必要となる．特にナレーションなどの音声のカット点は，発話者が一呼吸置くことによる無音区間を検出することによって対処することができると考えられる．

また，動きアクティビティについても，今回の実験では動きベクトルの大きさのみを用いているが，グローバルな動きとローカルな動きなど，動きベクトルの方向や分布も含めて考慮していく．同時に，動きアクティビティの閾値処理に必要なシーンの特徴判断についても検討する．

その他，反復ショットが存在する場合には代表ショットのみ抽出するなど，今後，このような改良について検討を行うと同時に，様々なジャンルのオーディオビデオコンテンツについて検証を行うこととする．

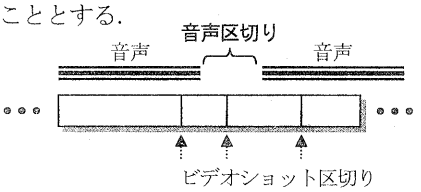


図 7　オーディオとビデオの非同期性

5.　まとめ

MPEG で圧縮されたオーディオビデオデータから，その内容を短時間で把握するために必要な自動要約抽出方式について検討した．本方式は，圧縮データ上でオーディオやビデオの時空間的な特性を解析することにより，抽出処理を自動且つ低コストに実現する．また，要約情報はコンテンツから均一的に抽出されるため，全体の内容把握を容易に行うことができる．MPEG-1 のデータを用いた予備実験により，意味的に重要な内容を持つ要約情報を構成することができた．

謝辞

日頃ご指導頂く KDD 研究所の秋葉所長，並びに浅見副所長，松島副所長に感謝致します．

参考文献

[1] ISO/IEC CD 15938-5, Information Technology – Multimedia Content Description Interface – Part 5: Multimedia Description Schemes, 2000

[2] 菅野，中島，柳原，「圧縮 AV データからの自動要約抽出に関する検討」，2001 年信学総大，2001（掲載予定）

[3] 田中，脇本，神田，「シーン検出による動画情報の自動要約・閲覧技術の開発」，信学技報 IE99-20, 1999

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

[5] R. Lienhart, et al., "Video Abstracting", Communications of ACM, Vol. 40, No. 12, 1997

[6] 中島，「フレーム間輝度差分と色差相関による圧縮動画像データからのカット検出」，1994 年信学春大，1994

[7] Y. Nakajima, Y. Lu, et al., "A Fast Audio Classification from MPEG Coded Data", SPIE ICASSP99, 1999