情 報 メ デ ィ ア 32－6
（1997. 9. 12）

映画におけるシーンの抽出を利用した階層的なビデオブラウザの構築

村野井 亮治　早坂 里奈　太田 浩二　趙 継英　松下 温
慶應義塾大学 理工学部
{muranoi,hayasaka,ohta,zhao,on}@myo.inst.keio.ac.jp

動画像の概要把握や検索を行ないやすくするビデオブラウザの実現を目的とし、同じ場所で起こった時間的に連続した場面をシーンと定義して、動画像からシーンを自動的に抽出するシステムを提案する。本システムは、カット点検出による動画像のショットへの分割、ショットの内容を表現する代表フレームの選出、ショット同士の類似度を用いて似ているショットをまとめることによるシーンの抽出からなる。さらに、抽出されたシーンを利用した階層的なビデオブラウザを構築する。

Hierarchal Video Browser based on Scene Extraction for Movie

Ryoji Muranoi , Rina Hayasaka , Koji Ohta , Jiying Zhao , and Yutaka Matsushita
Faculty of Science and Technology, Keio University

This paper presents a novel system that automatically extracts scenes from MPEG video, and a very user-friendly video browser. We define scene as the coherence consisting of shots and represents a sequence of events happened in the same place. The proposed system first divide video into shots by detecting cuts and chooses the representative frames of shot and extracts scene by synthesizing similar shots based on the similarity of shots. Then we build the hierarchal video browser based on scene extraction.

1 はじめに

近年、DVDなどのディジタル蓄積メディアや多チャンネルTVなどの新しい放送メディア、また、MPEGに代表される符号化技術の発達により、動画像を利用できる環境が整いつつある。今後こうした新しいメディアの急速な普及が見込まれ、一般に利用できる動画像情報はさらに増大していくと思われる。

しかし、動画像の取り扱いは難しく、例えば、

- 動画像全体の概要を簡単に把握する
- 動画像中の希望するポイントに簡単にアクセスする

といった機能はビデオデッキのような高速再生または早送り・巻戻しといったもので実現されているが、速度に限界があるため効率が悪く、また人間の認識能力にも限界があるため正確さに欠ける。

このような問題を解決するために、カット点(カメラカット)によって区切られる部分(これを"ショット"という)を抽出し[1][2]、その代表フレームを表示するビデオブラウザ[3][4]が提案されている。しかし、通常ショットは数秒から数十秒の長さであり、これだけでは数時間の長さの動画像に比べて短過ぎ、その数が膨大になってしまう。

そこで本研究では、原則的に同じ場所で起こり時間的に連続した場面を"シーン"として定義し、ショットよりも大きなまとまりとして扱う。これにより動画像は、

シーケンス(動画像全体)
|
シーン
|
ショット
|
フレーム

という階層構造として捉えられる。本研究では、MPEG1で符号化された映画における、シーンの抽出を含むこのような階層構造を自動的に解析するシステムを提案する。また、解析された階層構造を利用して、映画を階層的に表現するビデオブラウザを構築する。

以下2章では動画像のショットへの分割について、3章ではショットの代表フレーム選出について、4章ではショットのシーンへの統合について述べ、5章では結果について、6章ではシーン抽出を利用したビデオブラウザについて述べる。

2 動画像のショットへの分割

2.1 MPEG1

まず、MPEG1の特徴について説明する。

2.1.1 画像タイプ

- Iピクチャ(Intra-coded picture)

  他のフレームの情報を使わずに(フレーム間予測を行なわずに)自身で符号化された画面。画面内のMB(マクロブロック)はすべてIMB(フレーム内符号化)である。

- Pピクチャ(Predictive-coded picture)

  過去のIまたはPピクチャからの順方向予測によってできる画面。画面内のMBはIMBとPMB(順方向予測)である。

- Bピクチャ(Bidirectionally predictive-coded picture)

  過去または未来のIまたはPピクチャからの双方向予測によってできる画面。画面内のMBはIMBとPMBとBMB(逆方向予測)とBiMB(双方向予測)である。

2.1.2 GOP構造

MPEG1では画像データは前後をもとにして作られているために、1つの画像だけでは完結した情報とはならない。よって何枚かの画像をひとまとまりにしたGOPを単位としてランダムアクセスを可能にしている。(図1)

図1: Group Of Picture (GOP)

GOP 内のピクチャ数やIまたはPピクチャの現れる周期に制限はないが、本研究では実用的に前者を15、後者を3とした。

2.2 カット点検出

ショットとは各カメラカット点の間の部分であり、このカット点を検出することにより、動画像のショットへの分割を行なう。本研究では、前節で述べたMPEGの特徴の中のBピクチャに着目してカット点検出を行なう。

Bピクチャには、IMB、FMB、BMB、BiMBがあり、図2に示すようにこのBピクチャのMBの構成を調べることでカット点を検出することができる。

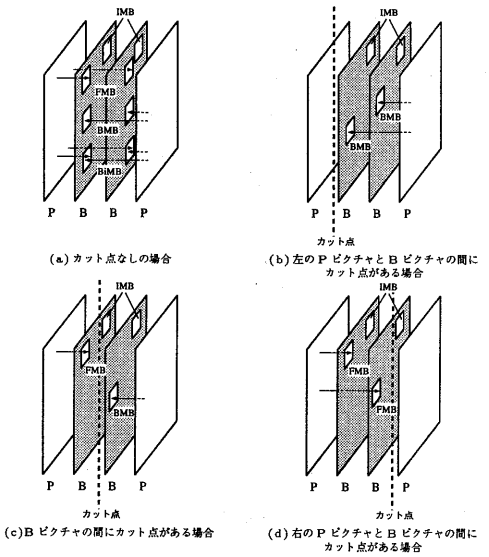

図 2: Bピクチャの MB 構成

カット点のない (a) では FMB と BMB がほぼ同じ数である。しかし、(b) は2つのBピクチャの FMB と BiMB が少なくなり、(c) は左のBピクチャの BMB と右のBピクチャの FMB と2つのBピクチャの BiMB が少なくなり、(d) は2つのBピクチャの BMB と BiMB が少なくなる。ここで以下の式で C_B を定義する。

$$C_B = \frac{N_F - N_B}{N_{Bi}} \qquad (1)$$

N_F、N_B、N_{Bi} は FMB、BMB、BiMB の数である。この C_B はカット点が存在するとその絶対値が大きくなる。したがって、この C_B を用いて以下の式よりカット点の有無を判断する。

$$|C_{B_1} \cdot C_{B_2}| > threshold_B \qquad (2)$$

さらに以下の式によりカット点の位置を調べる。

$$\begin{cases} P_1|B_1B_2P_2 & \text{if} \quad C_{B_1} < 0, \quad C_{B_2} < 0 \\ P_1B_1|B_2P_2 & \text{if} \quad C_{B_1} > 0, \quad C_{B_2} < 0 \\ P_1B_1B_2|P_2 & \text{if} \quad C_{B_1} > 0, \quad C_{B_2} > 0 \end{cases} \qquad (3)$$

ここで $P_1B_1B_2P_2$ はPピクチャ、Bピクチャである。

3 ショットの代表フレームの選出

ショット同士を比較する際に、ショットに含まれるフレームすべてを比較したのでは処理量が膨大になってしまう。またショットの内容を表示するためにもショットを代表するフレームを選出する必要がある。

本研究では処理量削減のためショット内のIピクチャのみを代表フレームの候補とし、Iピクチャを郡平均法を用いてクラスタリングすることにより、一つのショットに対し複数の代表フレームを選出することを可能とする。これにより長く動きのあるショットも内容を表現することができる。さらには仮にカット点検出に未検出が起こったとしても、複数の代表フレームを選出する本アルゴリズムにより、内容の欠落を防ぐことができる。

図3に代表フレーム選出のアルゴリズムを示す。具体的な処理手順は次のようになる。

1. ショット内のIピクチャを取り出す。
2. 動きの少ない部分を抽出し、これを初期クラスタとする。
3. 郡平均法によりクラスタリングする。
4. 各クラスタについてIピクチャを平均した平均画像を求める。
5. 平均画像と最も近いものIピクチャを代表フレームとする。

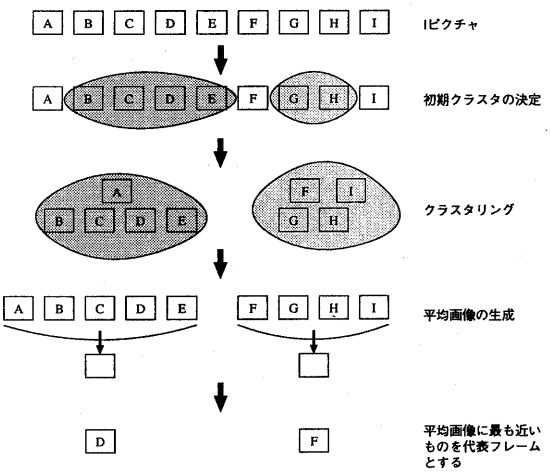

図 3: 代表フレーム選出のアルゴリズム

なお、I ピクチャの比較には I ピクチャを簡易復号化して用いる。この簡易復号化は、各 MB において色情報である DC 係数のみを復号化することにより、その MB の平均色が得られ、全体としてモザイク状のブロック画像 (DC 画像) が得られる。これによりさらに処理量を削減することができる。

4 ショットのシーンへの統合

たとえば会話のシーンなどでは、話者を交互に撮る場合が多いため、同じようなショットの繰り返しとなる。このように、ひとつのシーンの中には、似ているショットが含まれることが多い。この性質に着目してショットを統合し、シーンを抽出する。

4.1 ショット間の類似度

ショット間の類似度は、ショットの代表フレーム間の類似度を用いる。ただし、ひとつのショットが複数の代表フレームを持つ場合もあるため、すべての代表フレームの組み合わせについて類似度を調べ、最大値をショット間の類似度とする。

類似度はそれぞれの代表フレームにおける画素値の差と色ヒストグラムの差からファジー推論によって求められ、0〜1 の値をとる。

4.2 ショット間の結合度

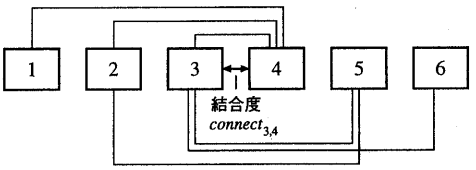

図 4: 結合度 ($N = 3$)

ショット $shot_n$ とショット $shot_{n+1}$ が連続している度合を表す結合度 $connect_{n,n+1}$ を、まわりのショットの類似度 s_{ij} を用いて以下のように求める。

$$connect_{n,n+1} = 1 - \prod_{i=n-N+1}^{n} \prod_{j=n+1}^{i+N} (1 - s_{ij}) \quad (4)$$

ここで N は比較するショットの範囲を表す。

このように、結合度 $connect_{n,n+1}$ はショット $shot_n$ とショット $shot_{n+1}$ だけでなく、その付近のすべてのショット間の類似度 s_{ij} から求められる。例えば図 4の場合、結合度 $connect_{3,4}$ は、類似度 s_{14}、s_{24}、s_{25}、s_{34}、s_{35}、s_{36} から求められ、ひとつでも類似度 1(そっくりなもの) があれば結合度は 1 になり、類似度 0 (似ていないもの) は結合度には影響せず、また類似度が 0.5(まあまあ似ているもの) でも柔軟に対応できる。

5 結果

5.1 対象動画像

本システムを実際の映画に適用した。用いた映画を表 1 に示す。

表 1: 対象動画像

	Length(M)	Total	Size	Source
A	31:52	45838	352 × 240	VideoCD
B	30:32	43942	352 × 240	VideoCD
C	30:09	43379	352 × 240	VideoCD

5.2 カット点検出

カット点検出の結果を表 2に示す。

ここで、Missは未検出数つまり検出できなかった数で、Overは誤検出つまりカット点ではないのにカット点としてしまった数である。

検出率はいづれも95％以上得られており、十分高い精度といえる。未検出の原因としては、人為的なディゾルブおよびノイズによりカット点が明確でなくなっている場合がほとんどであった。誤検出の原因としては、カメラおよび被写体の大きな動きによるものであった。

表 2: カット点検出結果

Video	Cuts	Detected	Miss	Over
A	272	270	2	11
B	377	359	18	0
C	266	253	14	0

5.3 シーンの抽出

式(4)によって求められた、結合度の変化を図5に示す。

これにより、変化の谷の部分をシーンチェンジとみなすことでシーンを抽出できる。シーンの抽出結果を表3に示す。

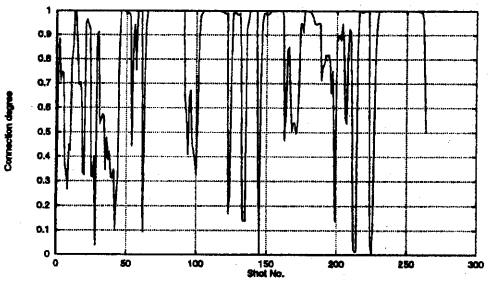

図 5: 結合度の変化

表 3: シーン抽出結果

Video	Scenes	Detected	Miss	Over
A	17	13	4	7
B	22	21	1	11
C	24	18	6	3

検出率はいづれも75％以上である。本手法が意味解析や知識を用いていないことと、代表フレームを複数選出して表示できることから考えると十分実用的であるといえる。

6 階層的なビデオブラウザ

本システムの解析結果を利用して、シーンおよびショットを階層的に表示するようにしたビデオブラウザを図6に示す。

図中左上のパネルは、従来のビデオデッキ同様の機能を提供するものであり、またモニタでもある。

右側のウィンドウは、シーンを表示するシーンウィンドウである。ここではひとつのシーンをひとつのアイコン(代表フレーム)で表している。ユーザは詳細を知りたいシーンをクリックすることで、そのシーンに含まれるショットを見ることができる。

左下のウィンドウがそのショットウィンドウである。このショットウィンドウは複数表示することが可能で、このとき選択されたシーンは色枠で囲まれて、対応するショットウィンドウも同色で表示されるので、対応関係を容易に把握することができる。

シーンウィンドウ、ショットウィンドウのいずれにおいても、アイコンをクリックすることでそのシーンまたはショットを再生することができる。

さらに、シーンはひとつのアイコンで表現することが困難な場合があるので、このブラウザではユーザの要求によって、シーンを表現するアイコンの数を以下の3段階に切替えることもできる。

- 1つのアイコンで表す

  ショットの代表フレーム選出と同様にして、各ショットの代表フレームをクラスタリングし、各シーンの代表フレームを選出する。このうち最大のクラスタから選ばれた代表フレームを表示する。

- 3つのアイコンで表す

  上記のシーンの代表フレームのうち、要素数が多い順に3つ、時間順に表示する。

- 代表フレームすべてを表示する

  すべてのシーンの代表フレームを時間順に表示する。

図 6: 階層的なビデオブラウザ ("Stand by Me"(©1986 Columbia Pictures Entertainment Inc.))

このようなブラウザによって、おおまかな情報を得てから必要に応じて詳細を把握したり、あるいは再生を行なうといったことが可能となる。

7 おわりに

本研究では、MPEG1 で符号化された映画を自動的に解析して階層構造を得る手法を提案し、それをビデオブラウザに適用した。本システムは、高速性と意味解析を行なっていないこと、さらには実際のブラウザにおける「動画像全体の概要把握」や「動画像中の希望するポイントへのアクセス」といったことへの操作性を考慮すると、十分実用的であるといえる。

今後の課題としては、さらなる精度の向上があげられる。しかし、画像処理で知識を与えずに意味解析を行なうことは困難であるため、別の特徴を利用した手法との組み合わせや音声解析との組み合わせといったことが考えられる。

本システムによるビデオブラウザはブラウジングだけでなく、編集、インデックス作成など動画像におけるさまざまな操作に利用可能であり、今後の動画像情報の効率的な処理の基盤となることが期待される。

参考文献

[1] H.J.Zhang, C.Y.Low, Y.Gong, and S.W.Smoliar. Video parsing using compressed data. In *Proceedings of the SPIE - The International Society for Optical Engineering*, Vol. 2182, pp. 142–149, 1995.

[2] Y.Taniguchi, Y.Tonomura, and H.Hamada. A method for detecting shot changes and its application to access interfaces to video. pp. 538–546, 1996.

[3] M.Shibata. A description model of video content and its application for video structuring. In *Transactions of IEICE*, pp. 754–764, 1995.

[4] Y.Taniguchi, A.Akutsu, Y.Tonomura, and H.Hamada. An intuitive and efficient access interface to real-time incoming video based on automatioc indexing. pp. 25–33, 1995.