ヒューマンインタフェース 67－7
（1996．7．11）

映像ブラウジングのための類似ショット統合

青木恒　下辻成佳　堀修
(株) 東芝 研究開発センター 情報・通信システム研究所
〒210 川崎市幸区小向東芝町1
{aoki,shimotsuji,hori}@eel.rdc.toshiba.co.jp

映像内容をブラウジングするためにカット点ごとに代表フレームを生成する方法が提案されている．しかし，映画などの長時間映像に関しては代表フレームが大量に生成されるという欠点がある．一方，映画などでは同じ場面の中で2つ以上のカメラを繰り返し切り換えて描く手法が多用されている．この典型的な例が対話シーンである．対話シーンでは，人物ごとのカメラアングルで撮影された映像が繰り返し出現する．本研究報告では，カメラアングルが同一と判定されるショットをグルーピングし，カメラアングルが異なるショットだけから代表フレームを生成することによって，代表フレーム数を従来手法の60%に削減した．また，その結果を利用した視認性の良い映像ブラウザを報告する．

A Shot Classification Method to Select Effective Key-frames for Video Browsing

H.Aoki, S.Shimotsuji, and O.Hori
R & D Center, TOSHIBA Corporation
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, Kanagawa 210, Japan

Recently, many methods to list key-frames representing the contents of video using cut-detection technique have been proposed. However, only by the cut-detection, a large number of key-frames are extracted from long sequence such as movie or drama. On the other hand, movies have many repetitous shots, mostly in dialogue scenes. In this paper, we describe the function of our video browser, which selects and shows only effective key-frames. The browser reduces about 40% of the key-frames and provides efficient user-interfaces.

1 はじめに

CD-ROMの普及やDVDの登場などにより，映像メディアに対するランダムアクセスが高速かつ容易になりつつある．これらのメディアを通じて，多くの映画やドラマといった長時間の映像をブラウジングし，すばやく内容を把握したいというニーズが高まると予想される．

映像内容を速読するためには，カット点を自動検出し，カットごとの静止画を一覧表示する方法が多く提案されている[長坂96][谷口96]．カット検出法およびその応用例も数多く提案され，実験が試みられている[Otsuji94][Zhang93][Zabih95]．

しかしながら筆者らの調査の結果，映画の場合，カットは約5秒に1回の頻度で現れる．したがって，カットを代表フレーム生成の単位とすると，2時間の映画に対して1,440枚の代表フレームが生成される．複数のカットから上位層を生成し，階層的な代表フレーム提示を行うこと[Zhang95]が，ユーザインタフェースの観点から必要である．しかし，人間がシーケンスに対して抱いている階層構造（「幕」「場」のような）に即した映像の階層化を行うためには，高度な内容理解が必要になる．

とりわけ映画やドラマにおいては，同じ場面内に複数のカメラを設置し，それぞれのカメラからの映像が交互に，繰り返して写し出される手法が多く用いられている．その最も典型的な例が複数の人物の対話シーンである．映画50分ぶんに対して調べた結果，43分間(87%)が人物登場シーンであり，そのうちの39分間(43分に対する89%)が対話区間であった．対話シーンにおいては，対話に参加している人物を撮影した映像がくりかえしながら交互に出現する．(図1はカットごとの代表フレームを列挙した例であるが，上から2番目と4番目のショットは同一カメラアングルから撮影された同一人物の映像である．同様に5番目と7番目と9番目，また6番目と8番目も同一アングルである．）筆者らはこのことに着目し，平易な類似画像判定法と判定の安定度を高める手法を用いて映像中のショットの繰り返し部分の自動検出を行った．そしてその繰り返しを省いたアイコン表示により，より少ない数の静止画一覧で映像内容全体を把握できるブラウジング・システムを開発した．

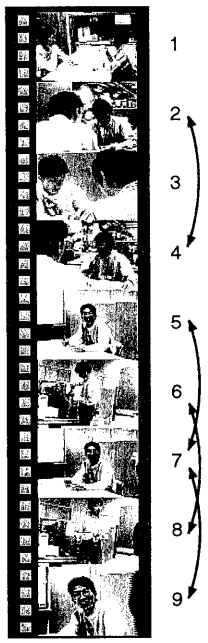

図1: 繰り返しショットの例

以下，第2章で類似ショット統合手法について説明し，第3章でその統合性能を評価し，処理結果を用いた映像ブラウザを紹介する．最後に第4章においてシステム全体の問題点について論じる．

なお，本研究報告においては，映像の連続性がとぎれる時刻を「カット」，カットから次のカットまでの映像区間を「ショット」と呼ぶ．

2　類似ショット統合

筆者らが処理に用いた映像データは，ＭＰＥＧ２ビデオストリームを時間的，空間的に間引きしたものである．時間的にはＩピクチャのみを用いることにより，１秒あたり約２枚の画像に間引きした．また，ＭＰＥＧ２の画像内圧縮が8×8ピクセルの大きさの「ブロック」単位に Discrete Cosine Transform によって行われている [MPEG] ことから，各ブロックの直流成分だけを採用することにより，720×480ピクセルから90×60ピクセルに間引いた．これにより無圧縮画像で秒31MBのデータ量を秒32KBにまで削減し，計算速度を向上させた．

これとは別に，MPEG2ビデオストリームからカット検出を行う．処理は図2のような手順で行う．

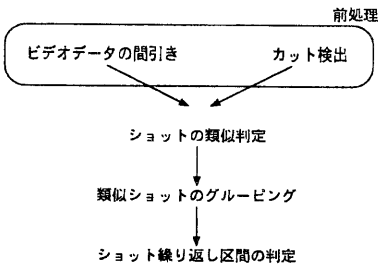

図 2: 処理の流れ

まず，検出された各ショットから数枚ずつフレームを取りだし，ショット間で類似のものがあるかどうかを検査する．異なるショットから取り出されたフレームで類似性があれば，２ショットは類似であると判定される．次に類似であると判定された２ショットの組合わせを結合していくことで，同一カメラアングルから撮影されたショットをグルーピングする．このグルーピングの際，類似であったショットの組合わせを複数比較することで安定な類似判定を確保する確認アルゴリズムを適用する．最後に，この結果を用いてショット繰り返し区間を検出する．以下ではこの処理手順を説明する．

2.1　ショットの類似判定

あるショットと，カメラアングルが同一であるショットを探すために，基準ショットと，基準ショットから数ショット後の対象ショットのそれぞれから静止画を取りだす．２静止画が類似かどうかを比較することによって，基準ショットと対象ショットが同一カメラアングルから撮影されたものであるかどうかを判定する．基準ショットから近傍の６ショット後までを探索する．

６ショットという数字は筆者らの実験によってヒューリスティックに決めた値である．

基準ショットと対象ショットから類似判定のためにとりだす静止画には，基準ショットの最終フレーム，対象ショットの開始フレームを選択する．これは，同一カメラアングルのショットは，間に別のショットが挿入された場合でも構図の連続性が保たれる傾向があるからである．つまり同一カメラアングルのショットが再登場する場合，一つ前の同一カメラアングルであるショットの最後の構図から映像が再開するということである．したがって，同一カメラアングルから撮影されたショットの組合わせであった場合，基準ショットの最終フレームと対象ショットの開始フレームが類似している．

本手法では，上記のようにまずショット端のフレームを比較し，それが類似であれば判定を中止，類似でなければ判定するフレームをショット内部に１つずつ進めていくという方法をとる．ショット内部に進めていくフレーム数は任意に決める，今回はショット端から３フレームまでの判定を行った場合を示す．

この３フレームという数字も実験によ

り，計算速度と性能のバランスが最適として導かれたものである．

2フレームの類似判定は画面全体の色相の類似と，画面各部の輝度分布の類似という2つの方法によって行う．2つの方法の双方で2フレームが類似であった場合に，基準ショットと対象ショットは類似であったと判定する．

画面全体の色相の比較には色相ヒストグラムを用いる．MPEG2では各画素は輝度Yと色相C_b, C_rで表されているが，色相に関して以下の極座標表示を行い，θ軸での度数分布をとったものである．

$$(C_b, C_r) = (\rho cos\theta, \rho sin\theta)$$

このヒストグラムは画面全体の色調を表現しているもので，同一カメラアングル内での輝度の変化やフェードに対して影響を受けにくい．ヒストグラムのθの分割を軸に，度数を座標にとると，フレームごとに超空間の1点として表現することができる（θを30分割したヒストグラムならば30次元）．この超空間内の距離がしきい値内であったものを，色相の上では類似と判定する．

一方，色相ヒストグラムだけでは画面全体としての色合いが似通っていれば，そこに写っているものが全く異なるものであっても類似と判定してしまう．そこで画面をブロック化し，2フレームで同じ場所にあるブロックの輝度差がしきい値以内であったブロックの数を計数することによって，画面レイアウトの類似性を検査する．ブロック比較の前には両フレームを画面全体の輝度の平均値，分散によって正規化する．これは色相ヒストグラムの際に述べたように，輝度の変化やフェードによる判定精度劣化を防ぐためである．

2.2 グルーピング

以上の過程で，類似であると判定された2ショットの組合わせを順次結合していくことによって，同一カメラアングルであったショットをグループ化する．

しかしながら，たとえば真っ黒な画面にフェードアウトしたショットと，真っ黒な画面からフェードインしたショットなど，本来は分断すべきショットの組合わせが，上記過程の類似判定で「類似」とされるなど，若干の不安定要因を残している（図3）．

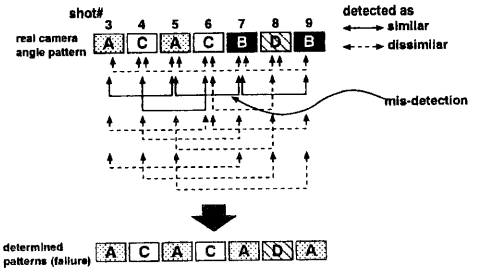

図 3: 類似ショット結合の不安定性

図3ではカメラアングル・パターンが異なるショット5とショット7を類似と判定しているが，これにより同図の下にあるようにカメラアングルBのショットはすべてカメラアングルAと同じグループに属してしまう．

ところで，ショット5とショット7の類似をみる際に，ショット9との類似を用いて再検討してみると，5と7は類似，7と9も類似であるのに，5と9が非類似であるというように矛盾が生じている．上述の例の場合，ショット5が黒にフェードアウト，ショット7が黒からフェードイン，そしてショット9がフェードなしの場合にこの状況になりうる．このように，ショットkとショットlとの類似判定について，別のショットmを含む組合わせで信頼性を評価

する.

ショット kとショット lの類似が変数 S_{kl} によって，類似の場合1，非類似の場合0と定義されていたとき，S_{kl}, S_{lm}, S_{km}からみた k,l 間の信頼性素因 $C_{kl(klm)}$を表1のように決める.

S_{kl}	S_{lm}	S_{km}	$C_{kl(klm)}$
1	1	1	1
1	1	0	0
1	0	1	0
0	1	1	0
1	0	0	1
0	1	0	−1
0	0	1	−1
0	0	0	−1

表 1: 信頼性素因の定義

こうして定義された信頼性素因を，とりうるすべての m に対して合計し，最終的なショット k,l間の信頼性 C_{kl}を以下のように決定する．C_{kl} は2ショットが同一カメラアングルであると判断すべきときに＋1，異なるカメラアングルと判定すべき時に－1であり，0に近いほど値の信頼性が低い.

$$C_{kl} = \left(\sum_m C_{kl(klm)} \right) / F$$

なお，Fは「とりうるすべての m」の個数である.

前出の例の場合，もしショット5－7間の類似判定が「非類似」であった場合には $C_{57} = -1$ となるが，図3のままの誤りであれば，この数値は0.6となり，信頼性が低い．たとえば $C_{kl} \geq 0.7$ などのしきい値を設ければ，信頼度の高い類似判定だけを選択的にグルーピングに寄与させることができる.

以上のようにして同一カメラアングルから撮影されたショットをグループ化したが，本システムではカメラアングル・グループの始まりから終わりまでを同一ショットの繰り返し区間として認識する．二つ以上のカメラアングル・グループがオーバーラップしている場合には，最初に登場したグループの始まりから，最後に登場したグループの終りまでが繰り返し区間になる．たとえば，ショットのカメラアングルパターンが「ABABABCABACBAB」である区間は一括して繰り返し区間と見なされる（最初のカメラアングル・グループ＝Aの始まりから最後のグループ＝Bまで）.

3 類似重複のない映像ブラウザ

図4には，従来の方法のようにショットごとに1枚の代表フレームを生成するブラウザの例を示す.

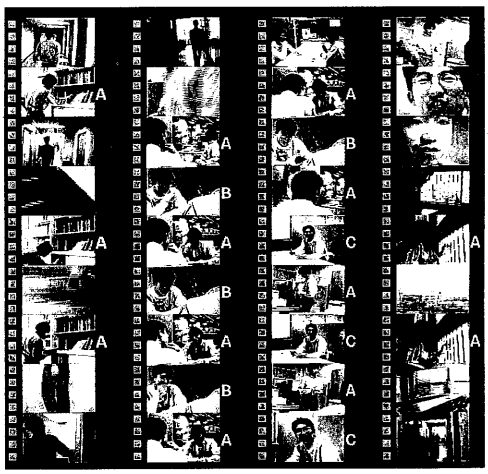

図 4: カメラアングル・グループの判定例

図 5: 繰り返し区間ごとの表示例

図 6: 類似重複ショットを省略した例

各アイコンの横に付与されているアルファベットは前節までで説明したカメラアングル・グループである．また，カメラアングル・グループのアルファベットが帯で結ばれているものは繰り返し区間として検知された部分である．

これを繰り返し区間，非繰り返し区間それぞれを横に並べて表示させると図5のようになる．また，繰り返し区間のうち同一カメラアングル・グループからは1枚ずつの代表フレームに絞ると図6のようになる．

次に，映画「Green Card」(Touchstone Pictures 1990年) の60分ぶんに対して本手法で代表フレーム省略を試みた結果を示す．実験対象映像は635ショットを含んでいるが，その中から59の繰り返し区間を検出した．類似重複ショットを省略して表示したところ，本来635の代表フレームを表示するところが381の代表フレーム表示になる（60%）ことがわかった．その一例として5分間の中に53ショットを含む対話シーンでは，システムは2つの繰り返し区間を検出し，代表フレーム数を13に減らすことができた（25%）．

またこの中で，誤ったカット検出によって分断された2つのショットを5ヶ所をすべて正しく再統合した．この誤ったカット検出とは，車など早い物体が画面全体を横ぎったり，クローズアップしたアルバムのページをめくったりしたことによるものである．

一方，今回の実験は映画を土台に行ったが，映画以外のジャンルの映像に対して本手法を適用した例とそれらの問題点を示す．ここでは，より実用性を示すためにカット検出技術を用いて定義されたショットを出発点とした．また，類似ショットを探索する距離は基準ショットから対象ショットまでの距離が最大6ショット，ショット内でフレームを比較する深さは3である．

●ニュース番組（1）
NHK Today's Japan
映像時間15分．検出ショット数121

同じカメラアングルのショットが再登場するのはキャスターのショットのみであった．しかし，ほとんどのニュース項目が設定値の6ショットより多くのショットを含んでおり，12項目のうち8項目までは，項目をまたいだグルーピングは行われなかった．残る4項目は，前のニュースのキャスター・ショットから次のニュースのキャスター・ショットまでを結んでおり，さらに先のニュースまでは結ばれていない．ニュース番組の場合には，基準ショットと対象ショットの距離を大きく取ってキャスターのショットを検出し，キャスター・ショットごとに1枚の代表フレームを表示させるのが最善策であろう．

一方，前半のニュース項目は国会の話題であったが，フラッシュがたかれたためにカット検出が誤ってなされ，2つに分断されたショットが16ヶ所あったが，これらはすべて類似ショットとして正しく統合された．また，答弁の部分1ヶ所（5ショット）は正しくグループ化された．結局代表フレーム数は79に削減された（65%）．

●ニュース番組（2）
NHK おはよう日本
映像時間10分．検出ショット数50

ニュース項目を越えてグループ化されたショットはなかった．また，カット検出の誤りもなかった．地方局からの中継のニュースがあり，東京のスタジオ，北海道のスタジオ，事故現場のショットが繰り返し登場したところでは，10ショット（1分1秒）を一つの繰り返し部分として検出した．最終的に代表フレーム数は36に削減された（72%）．

●ニュース番組（3）
東京メトロポリタンテレビ 新東京物語
映像時間10分．検出ショット数82

再帰画面が1ヶ所しか現れず，統合できたのは1ヶ所3ショットだけであった．このほかに，暗い画面でカット検出に失敗して過検出した部分を3ヶ所（2＋2＋15ショット）グループ化した．したがって代表フレーム削減効果は少なく，64にとどまった（78%）．

このように，対話部分を多く含む映画に対しては効果的に類似重複代表フレームを削減することができるが，それ以外のジャンルでも繰り返し構造を持つ映像については，削減効果があることがわかった．

4 まとめ

以上のように，映画に多く含まれる同カメラアングルショットの繰り返しを自動検知することによりカット検出だけで選択された代表フレームの数を半分程度にまで削減できることを示した．

今後の課題としては，

- カメラアングルが異なっていても同一人物が登場するショットをグループ化するために，写っているオブジェクト単位でのマッチングを行う必要がある．
- 映像ジャンルによって繰り返し登場するショットのパターンがあるが（たとえばニュース番組ならばキャスターの映像がニュース項目ごとに繰り返し現れる），ジャンル特有のパターン辞書を持つ必要がある．

という点が挙げられる．様々な映像ジャンルに対して適切なグルーピング手法を探っていくことも今後の課題である．

参考文献

[長坂96] 長坂晃朗，宮武孝文，上田唯博，“カットの時系列コーディングに基づく映像シーンの実時間識別法” 信学論(D-II),vol.J-79-D-II, no.4, pp.531-537, 1996.

[谷口96] 谷口行信，外村佳伸，浜田洋，“映像ショット切換え検出法とその映像アクセスインタフェースへの応用” 信学論 (D-II),vol.J-79-D-II, no.4, pp.538-546, 1996

[Otsuji94] K.Otsuji et al., “Projection-detecting filter for video cut detection,” Multimedia Systems. 1:205-210, 1994

[Zhang93] H.Zhang et al., “Automatic Partitioning for Full-Motion Video,” Multimedia Systems. 1:10-28, 1993

[Zabih95] R.Zabih et al., “A Feature-Based Algorithm for Detecting and Classfying Scene Breaks,” ACM Proceedings, pp.189-200. 1995

[Zhang95] H.Zhang et al., “Video Parsing, Retrieval and Browsing: An Integrated and Content-Based Solution,” ACM Proceedings, 1995

[MPEG] ISO/IEC 13818-2: “Information Technology — Generic Coding of Moving Pictures and Associated Audio: Video,” International Standard, 1995