

映像構造を利用した代表フレーム表示方法

青木 恒 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,hori}@eel.rdc.toshiba.co.jp

内容をユーザが把握するためのひとつの簡便なインタフェースとして代表フレーム一覧表示がある。このような手法では映像を場面の変わり目ごとに分割し、分割単位ごとに代表フレームを生成する方法がとられるが、長時間の映像では代表フレームが大量になるという問題点がある。本論文では映像が持っている構造に即して代表フレーム数を削減する方法を示し、それをういたユーザインタフェースについて議論する。

A Method of Displaying Representative Key-Frames Using Movie Structures

Hisashi AOKI and Osamu HORI

R&D Center, TOSHIBA Corporation

1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki Kanagawa 210, Japan

Key-frame listing is one of video browsing interfaces for quickly understanding the contents of movies. In such a method, video sequence is divided into shots by cut-detection technique, and a key-frame is extracted from each shot. However, a large number of key-frames may be produced from long movies. In this paper, the authors propose a method to reduce the number of key-frames, and also show video browsing interfaces employing this method.

1 はじめに

従来からのコンピュータ/インターネット上でのファイル利用に加え、最近では、DVD の発売やデジタル衛星放送の開始、デジタル・ビデオカメラの発売など、一般家庭へのデジタル映像導入は急速に進んでいる。

高画質、データサイズの高圧縮と並ぶデジタル映像の最大の特長に、ランダムアクセスがある。しかし映像(とくに映画など)の場合、音楽 CD における「曲」のような明確な切れ目がないために、ユーザのランダムアクセスが快適になるようなインデックス生成については未だ明快な手法が確立されていない。

自動化可能な映像分割の手法にカット検出がある。これは、映像の変わり目を自動検出して、変わり目ごとに

代表フレームを生成する方法であるが[1]、実際の映画の場合、15 分に 300 のカット点が存在するという報告[2]もあるように、こうして生成される代表フレームは数が膨大になるという問題点がある。

本論文が提案する代表フレーム表示方法では、カット検出によって分割された、映像単位の繰り返し構造を自動検出し、分割された映像単位を再統合[3]した代表フレーム表示を行うことによって、少ない代表フレーム数での内容把握を可能にする。また、本手法を適用したユーザインタフェースについても議論する。

2 手法の原理と利用

図 1左は、実際の映像をカット検出によって分割し、分割された小区間「ショット」ごとに 1 枚の代表フレームを

表示したものである。同図右には、それに対応して同じカメラアングルから撮影された映像に同じアルファベットを付した。ここに見られるように、特に会話シーン(図中の後半)においては、同じカメラアングルのショットB,C,Dが繰り返し現れている。

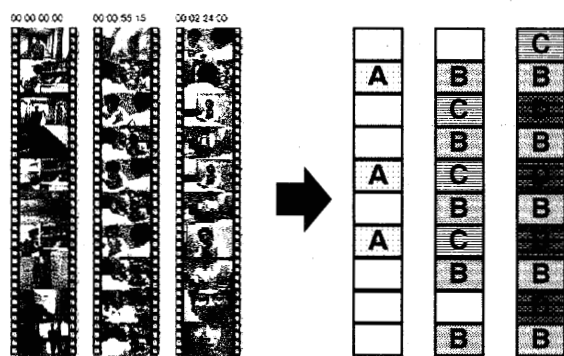


図1 代表フレーム一覧

実際の映像の「ひとつのシーン」には、このように同じカメラアングルから撮影されたショットが繰り返し登場する部分が含まれている。このように実際の映像の「ひとつのシーン」中で、繰り返しショットによって構成されている部分を「アクト(act=場)」と呼ぶことにする。

たとえば図2のようにカメラアングルパターンA,B,Cが出現していた場合、それぞれのパターンの出現範囲の論理和を持って「アクト」を検出することができる。(映画にもよるが)実際の映画では人物が登場している場面が多く、人物が登場している場面では対話が発生していることが多い。対話シーンは、このように「アクト」を検出することができる典型的な例である。著者らがある映画50分ぶんに関して調査したところ、そのうち43分間は人物が登場しているショットであり、39分が対話シーンであった[4]。また、対話シーン以外でも、例えば2つの場所で同時に起きている出来事を描く場合や、家に向かう人と家の中で待つ人を描く場面など、同一カメラアングルのショットが繰り返し登場するということは多く見られる。

本手法により類似ショット(同一カメラアングルから撮影されたショット)および、ストーリー単位「アクト」を自動検出することができる。これを利用して、カット検出によって生成された大量の代表フレームのうち、(1)同一カメラアングルから生成されたものの表示を省略する。(2)「アクト」の先頭ショットのキーフレームのみを表示する、という2

種類のユーザインターフェースを作成した。

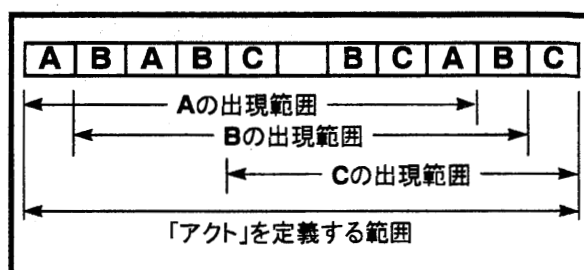


図2 「アクト」の定義

3 カメラアングル・パターンの検出

3.1 前処理

処理には、ディジタル圧縮の標準形式であるMPEG2ビデオ・ストリームを用いた。各フレームは720×480ピクセルで構成されており、秒あたり30フレームが含まれている。この形式はDVDなどで使用されている典型的なデータフォーマットである。映像はMPEG2ファイルに含まれる動きベクトルを利用して高速カット検出される[5]。

一方、類似ショット探索速度向上のため、映像データから解像度と時間を間引いた画像を作成し、それを用いてショットのグルーピングを行う[6]。著者らが用いたMPEG2ビデオデータは15フレーム(0.5秒)ごとにIピクチャを含む圧縮が施されたものである。Iピクチャだけを処理対象とした。また、各Iピクチャは大きさ8×8ピクセルのブロックを横に90個、縦に60個並べて離散コサイン変換されてコード化されているので[7]、各ブロックの直流成分だけを取り出すことによって、90×60のサイズに間引いた画像を作成する。以下では便宜上、こうして生成された大きさ90×60ピクセルの画像のことを単に「フレーム」という。

3.2 ショット比較のためのフレーム選択

上記によって映像がショットに分割された。次に、同一のカメラアングルであったショットを探索するが、これは各ショットから数枚のフレームを取り出し、それらの類似度を比較することによって行う。本来ならばショット中全てのフレームに対して類似比較を行うべきであるが、このようにすると例えば5秒と10秒のショットを検査する際には、

$(5 \times 2) \times (10 \times 2) = 200$ 通りの類似計算を行わなくてはならず、ショット数が増えた場合には組み合わせの数は莫大になる。

実際の映像では、同じカメラアングルから撮影されたショットの場合、ショットが推移していても、ほとんどのケースでは前ショットが終了した状態で次のショットが再開されていることを利用し、ショットの終端からの数フレーム、およびショットの開始点からの数フレームに関してのみ、類似計算を行う。

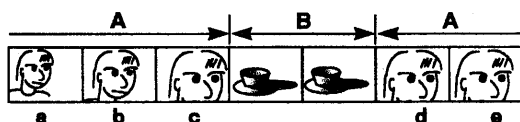


図3 継続するショットのフレーム

たとえば、比較するフレーム数を 2 にした場合、図 3 では、**b-d**, **b-e**, **c-d**, **c-e** の 4 通りの組み合わせについてのみ類似計算を行う。この例の場合、実際にはショットの終端 **c** から **b**, **a** と離れていくに従い、**d** および **e** との類似は失われてゆく。したがって、このように比較範囲を限定しても類似検査には大きな悪影響は及ぼさないと考えられる。実際に何フレームをショット端から採用するかについては実験的に決定し、今回の実験では、3 フレームという数字を選んでいる。したがって、2 ショットの類似を比較するために用いられるフレームの組み合わせは $3 \times 3 = 9$ 通りである。

3.3 ショット探索範囲

同一カメラアングルで撮影されたショットを探索する範囲も計算速度に大きな影響を及ぼす。例えば 2 時間、900 ショットの映画を処理するにあたって、全てのショットの組み合わせを計算したとすると、 $900C_2 = 40$ 万通りのショットの対に対して類似か否かを検討しなければならないが、実際には 2 時間の最初のショットと最後のショットが同じカメラアングルであることは少なく、また、仮にそうであったとしてもそれらを一つの「アクト」に統合してしまうことは意味がない。従って類似のショットを探索するためには、あるショットに対してその後数ショットに限って行えばよい。

ショット数が N のとき、全てのショットの組み合わせを比較した場合、組み合わせの数は $\frac{1}{2}N(N-1)$ である。それに

対して n ショットに限って比較を行った場合、組み合わせの数は $\left(N - \frac{1}{2} - \frac{1}{2}n\right)n$ である。900 ショットに対して、ショットの探索範囲を 8 ショットとした場合、7,164 通りの比較だけで済む。

具体的な探索ショット数の決定も実験調査に基づいて行った。著者らが、あるショットに対して、同一カメラアングルのショットが何ショット後に再登場するかという調査を 4 種類の映画 44 分 40 秒に対して行ったところ、その 96% は 8 ショット以内に現れていた。これに基づき、今回の実験ではショットの探索範囲は基準ショットから 8 ショットまでとした。

3.4 フレームの類似比較

フレームの類似度は、色成分と輝度成分の両方を用いて相補的に行う。

3.4.1 色成分

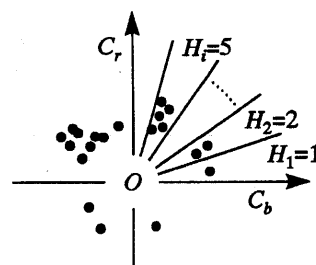


図4 色相ベクトルの定義

MPEG による画像圧縮では、各ピクセルは輝度 Y ($0 \leq Y \leq 1$) および色差 C_b, C_r ($-0.5 \leq C_b, C_r \leq +0.5$) によって表現されている。このうち色差成分だけを極座標変換し、 $(C_b, C_r) = (\rho \cos \theta, \rho \sin \theta)$ と表すと、 θ 方向のヒストグラムを作成することができる。すなわち、 $\frac{2(i-1)}{N}\pi \leq \theta < \frac{2i}{N}\pi$ の範囲にあるピクセル数を H_i とする(図 4)。

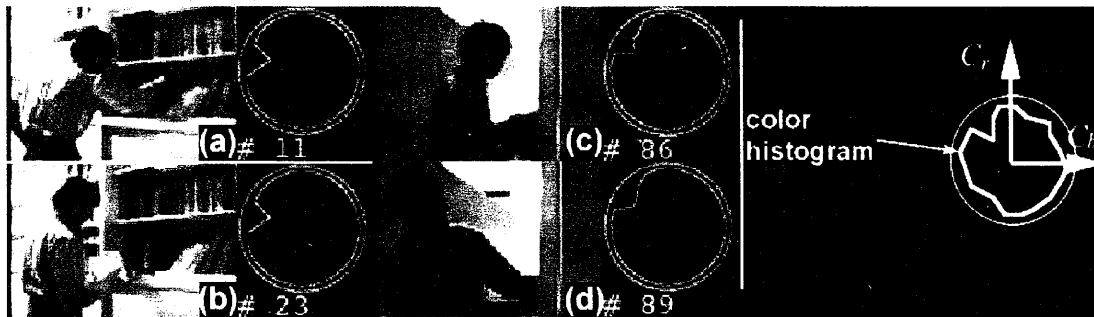


図5 色相ヒストグラム

このように計算される色相ヒストグラムから、フレーム全体の色成分の特徴量を表す色相ベクトル $H=(H_1, H_2, \dots, H_N)^T$ を定義することができる。2つの時刻 t_1, t_2 のフレームにおける色相ベクトル $H(t_1), H(t_2)$ の非類似度は、2つのベクトルの2乗距離 $d=|H(t_1)-H(t_2)|^2$ で計算する。なお、 ρ の小さなピクセル (グレイスケールに近いピクセル) に関しては θ 計算の誤差が大きいためにヒストグラム作成から除外する。

図5には、実際の映像から生成した色相ヒストグラムの例を示す。図の(a),(b),(c),(d)は、それぞれ別のショットに属するフレームである。図からわかるように、同じカメラアングルから撮影されたフレーム(a),(b)および(c),(d)は、オブジェクト(登場人物)の移動にも関わらずヒストグラムの形状が類似であるが、カメラアングルが異なると形状に違いが見られる。また、図には示されていないが、同じカメラアングルの画像であっても、フェイドの画像効果が施されている場合、あるいは太陽が照ったり曇ったりしている場合のように、画像全体の輝度が変化しているときにも、この色相ヒストグラムは安定な形状をとる。

3.4.2 輝度成分

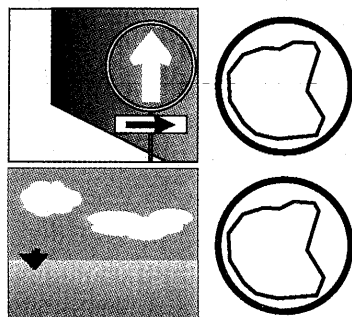


図6 異なる画面で似た色調

一方、図6に示したように、フレームの内容は明らかに異なるのに、全体としての色調が同じなために、前節の色相ヒストグラムの比較においては「類似」と判定される場合がある。この誤判定は、画面全体が暗いなど、コントラストが低い場合にも生じる。前節での ρ の値が小さく、ヒストグラムのカウントからのぞかれるためである。

このような誤判定を避けるために、色差成分を除いた後の輝度成分を用いた類似判定を並行して行う。

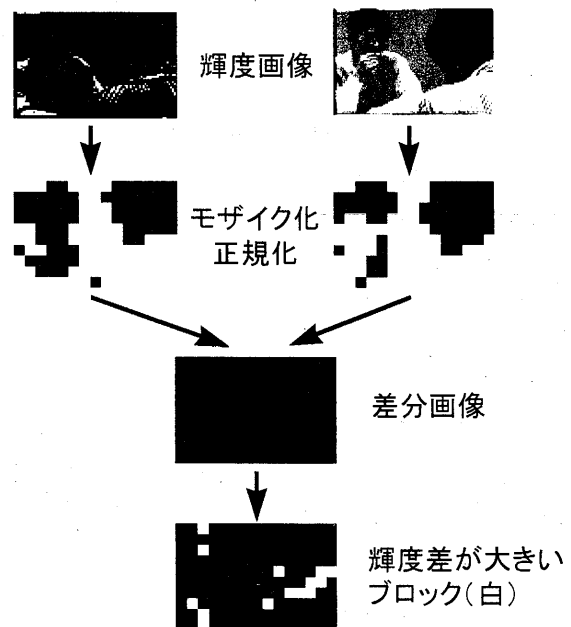


図7 輝度成分を用いた(非)類似度計算

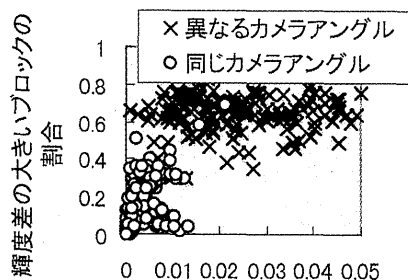
図7を参照しながら手順を説明する。まず前処理として画像のモザイク化および正規化を行う。モザイク化は画像内オブジェクトの微動の影響を押さえるため、 90×60 ピクセルのフレームを 15×10 個のブロックに分割し、各ブロック内では輝度値の平均を行う。全てのフレームに

関して全ブロックの輝度の平均値と分散が等しくなるように正規化を行う。この正規化は、この章の初めに述べたように、コントラストが画像によって異なる影響を抑えるために行う。

次に、モザイク化、正規化を行った画像同士を比較し、輝度差の絶対値を求める。その輝度差の値が実験的に定めたしきい値よりも大きかったブロックの数をカウントし、それが全ブロック数に対して占める割合をもって 2 画像の非類似度(距離)とする。

3.4.3 色差・輝度の併用による弁別性能

以下では、実際の実験結果に基づいて、色差と輝度の併用による効果を示す。



色相ベクトルの2乗距離
図8 併用による効果

図8は、4種類の映画(合計44分40秒)から無作為に選んだショットの組3,666組に対して、説明した色差および輝度成分による類似比較を行った結果である。どちらか一方だけを用いたクラス分けは難しいが、2類似度の併用によって、それらが同じカメラアングルから撮影されたものか否かをより正確に求めることができる。著者らは、この実験結果に基づいて色相ヒストグラムの距離のしきい値を0.026、全ブロックに対する輝度差の大きなブロックの割合のしきい値を48%と定めた。このとき「同一カメラアングルの未検出率」は2.0%、「異なるカメラアングルを誤って同一と判定する率」は1.2%であった。

3.5 アクト定義

3.5.1 類似度判定の検証

ここまでの手順で同一カメラアングルから撮影されたショットを検出することができたが、実際には2種類の誤り

が生じうることを前節の最後に述べた。すなわち「同一カメラアングルであるにもかかわらず見落とす(false negatives. 以下、判定漏れという)」「異なるカメラアングルを誤って同一と判定する(false positives. 以下、過検出という)」である。両者ともに低いことが手法としては最善であることは言うまでもないが、同じ誤りでも後者はより減らさなければならないものである。

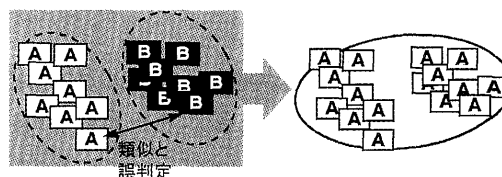


図9 過検出に起因する誤ったグルーピング

なぜなら、過検出により、実際には無関係なショットが「同一カメラアングルである」と判定されてしまうことは、2つ(あるいはそれ以上)の無関係な場面を1つとして統合してしまうおそれがあるからである(図9)。実際には1つのカメラアングルであったショットが、2つのショット群として判定された(判定漏れ)場合には、ユーザは2つのグループを指定し「これらを統合せよ」と指示することで修正が可能であるが、過検出によってグループ化されてしまった1つのショット群を2つに分断するためには、その一方の群に属するショット全てを指示しなければならず、ユーザの負担は大きい。

図10には2種類の誤りに対する修正の手間を図示した。過検出によって統合された2つのショット群のうち、少ない群の方のショット数が n であったとき、修正の手間は $(n+1)$ 回に及ぶ。

そこで、以下の手法によって、前節までのショット類似判定の正確さを検証する。

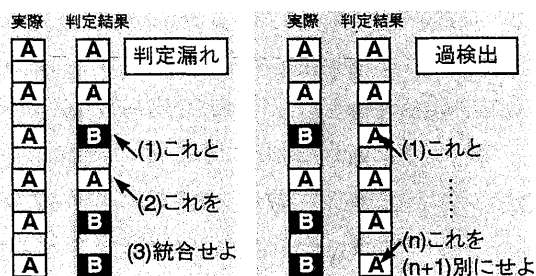
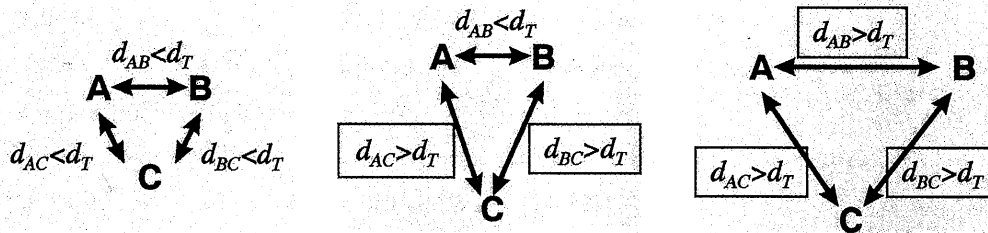


図10 誤りの種類と修正の手間



2つのショット **A**, **B** 間の非類似度 (距離) を d_{AB} と表すことにする。 **A**, **B** と別の参照ショット **C** に関する非類似度 d_{AC} , d_{BC} を導入すると, d_{AB} , d_{AC} , d_{BC} を 3 辺とする三角形を想定することができる。さて, 前節までのショット類似判定の結果は距離のしきい値 d_T よりも d が小さいときに「類似」と判定されたものと仮定すると, ショット **AB** 間, **AC** 間, **BC** 間の判定結果が次の 3 つのケースである場合には, d_{AB} , d_{AC} , d_{BC} は, 類似・非類似条件 $d < d_T$ または $d > d_T$ を満たしながら, いかなる数字もとることができる。すなわち 3 つのショットが「類似」「類似」「類似」または, 「類似」「非類似」「非類似」または「非類似」「非類似」「非類似」である場合である (図 11)。

しかし、例えば 3 つのショットの関係が「**AC** は類似」「**BC** は類似」「**AB** は非類似」である場合、は非類似条件 $d_{AB} > d_T$ の他に $d_{AB} < d_{AC} + d_{BC}$ に拘束され、 $d_T < d_{AB} < d_{AC} + d_{BC}$ の範囲しか値をとることができない。 d_{AC} , d_{BC} の値によっては、とりうる範囲がない場合さえある。著者らは前節までに述べたように色相ベクトルの 2 乗距離と、輝度の差が大きいブロック数という尺度の違う 2 つの基準によって類似か非類似かだけを判定している。このため具体的な d の数字を決定することはできないので、 $d_T < d_{AB} < d_{AC} + d_{BC}$ という拘束がある場合に、仮想的な d_{AB} , d_{AC} , d_{BC} が矛盾がないものであると保証することができない。したがって、前節までの 3 ショットの類似判定結果が「類似」「類似」「非類似」であった場合には、これら 3 組の類似判定のうちいずれかに誤りの恐れがある。

実際の計算では、類似度計算を行った全てのショットの組み合わせから 3 つずつの組を作成し、それらが「類似」「類似」「非類似」である場合には何もせず、それ以外の場合には「類似」であった組に 1 点を加点、「非類似」

であった組に1点を減点する。このようにして全ての3ショットの組み合わせの合計を計算すると、大きな正の値を持つショットの組は同一カメラアングルである確信度が高く、大きな負の値を持つショットの組は異なるカメラアングルから撮影された確信度が高い。前節までの類似度判定に誤りが多く含まれている場合にはこの値は0に近くなる。

3.5.2 グループ化

以上のように、「高い確信度」で同一カメラアングルから撮影されたと類推されるショットの組を選択することができるので、これらに同一カメラアングルのパターンIDを振ってゆく。パターンIDが決定した後には、第2章で説明した手法で「アクト」をラベル付けてゆく。

図 12には以上により判定されたカメラアングル・パターンおよびアクトが示されている。カット検出によって抽出された代表フレームの横にはアクト番号, そしてアクト中でのカメラアングル・パターン ID がアルファベットで示してある。



図12 映像構造判定例

4 実験結果とユーザインタフェース

著者らは実際の映画 120 分ぶんを用いて代表フレーム数削減の効果を実験した。なお、実験に用いた映画は、前章までで各種しきい値を定めるために用いた映画とは異なるものである。表 1には、実験に用いた映画と、それに関する諸量を示す。

映画名	時間(分)	ショット数
グリーン・カード	60	635
ホーム・アローン 2	30	444
シャーロック・ホームズ	30	283
合計	120	1,362

表1 実験に用いた映画

これらの映画に関しては、同一カメラアングルから撮影されたショットの組の正解を手入力した。この正解と比較して、本手法による検出結果は、判定漏れの率が 22.7%，過検出の率が 5.9%であった。ここでの数字は過検出率を最小にするようにしきい値を決定したものであることを特記しておく。

上記の手法を利用して同一カメラアングルから生成された代表フレームを省略する映像内容一覧インタフェースを 3 種類作成した。一つは、単に同一カメラアングルに対して 1 枚の代表フレームを表示するもの、残りの 2 つは、前章で自動検出した「アクト」の単位で代表フレームを表示するものである。表 3には、それぞれの表示方法および長所・短所を挙げた。タイプ 2 とタイプ 3 との違いは、アクトに属さない「谷間」ショットの取り扱いについてである。自動アクト検出では、実際の場面の終わり近くや導入部など、繰り返し同じカメラアングルが現れない場合には、アクトに属さないショットが生じる。このようにして自動検出されたアクトの間にある、アクトに属さないショットのことを「谷間」ショットと表記している。タイプ 2 では谷間ショットをそのまま表示しているために、アクトを表すアイコン 1 つと複数の谷間ショットの一つ一つが等価になってしまうという短所がある。図 13には、それぞれのタイプの表示例、表 2には表示タイプ別のアイコン数を示した。

この実験の結果、全体で 71.8%のショットをアクトに属させることができた。実際の場面の中に複数のアクトを定義するため、検出されたアクト数は、実際の場面数の 1.73 倍であった。それぞれの実場面内では平均 57.1%

のショットがアクトに属している。また、タイプ 3 のインタフェースにおいては、複数の実場面にまたがる谷間ショットを一つの代表フレームにしてしまった率は 33.3%，複数の実場面にまたがってアクト定義をしてしまった率は 3.0%であった。

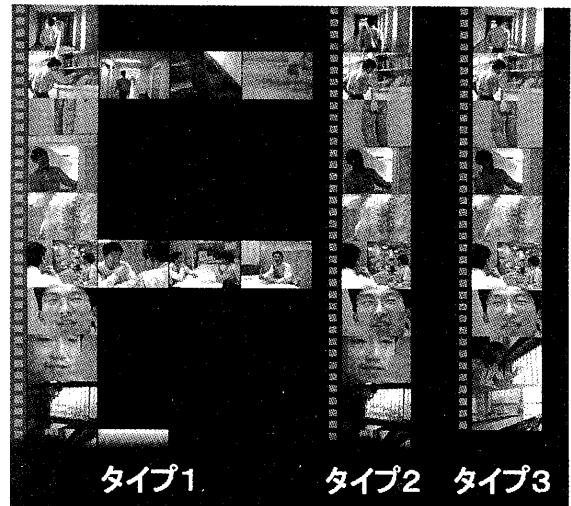


図13 ユーザインタフェースの表示例

映画	検出アクト数	表示した代表フレーム数 (削減率%)		
		タイプ 1	タイプ 2	タイプ 3
グリーン	58	284 (55.3)	186 (70.7)	97 (84.7)
アローン	35	186 (55.9)	109 (75.5)	60 (86.5)
ホームズ	18	119 (58.0)	56 (80.2)	31 (89.0)
合計	109	599 (56.0)	351 (74.2)	188 (86.2)

表2 インタフェースごとの代表フレーム表示数

著者らが製作したインタフェースでは、タイプ 2、タイプ 3 において代表フレームの縮約による情報の欠落を防ぐために、ユーザがマウスポインタをアクトの代表フレームに置くと、その場でそのアクトの内容が再生されるように設計されている。また、タイプ 3 においては、1 枚の代表フレームに縮約された「谷間」ショットに関しても、同様にマウスを目的の代表フレーム上に置くことで内容確認再生ができるようになっている。

種類	アクト検出	表示方法	長所	短所
タイプ 1	利用しない	カメラアングルごとに1枚の代表フレーム	代表フレームの削減は最小限なので、情報が失われない	削減効果が少ない
タイプ 2	利用する	アクトは1枚の代表フレーム	ストーリー単位に即した省略を実現	「谷間」ショット1つと、アクト1つが面積上等価になる
タイプ 3		アクト間の「谷間」ショットも1枚の代表フレーム	最小限の代表フレーム数で映像内容が把握できる	「谷間」ショットが実際の2場面にまたがる恐れがある

表3 ユーザインタフェース

5 まとめ

以上述べたように、本論文の手法によると、色相ベクトルの2乗距離、および、輝度をモザイク化・正規化した画像中での輝度差の大きいブロック数をカウントする方法の併用により、映像中で同じカメラアングルから撮影されたショットを検出することができ、また、そのようにして検出された3つのショットの類似・非類似関係を利用して、その検出精度をより高めることができた。「同じカメラアングルから撮影されたショットに関しては代表フレームを一つにする」という内容一覧表示を行うと、カット検出によって生成された代表フレーム画像の56%を削減でき、また、同一カメラアングルの繰り返し出現を検出することで場面のまとめ「アクト」を自動検出し、アクトごとに代表フレームをまとめることによって、最大で86%の代表フレーム量を削減できることがわかった。

さらに、アクト単位で代表フレームを統合することによる表示情報の不足を補うために、統合された代表フレーム表示においては、代表フレーム自身が動画像となって内容表示を行うことによってアクト内容を表示するインタフェースを製作した。

今後の課題としては、本手法を用いたユーザインタフェースを実際に被験者に使ってもらい、使いやすさを定量的に検証する実験を行うことの他に、映像内容により自動的に各種パラメータを設定する方法を検討することや、「谷間」ショットの統合や表示をどう行うかという問題を解決することが挙げられる。また、ニュースなど既知の構造を持った映像に適用し、同一カメラアングル(=ニュースキャスターのショット)を自動的に検出し、ニュース項目ごとにインデックス付けを行う方法などについて取り組む予定である。

参考文献

- [1] 上田博唯, 宮武孝文, 吉澤聡「認識技術を応用した対話型映像編集方式の提案」信学論(D-II), vol.J-75-D-II, no.2, pp.216-225, 1992.
- [2] M.M. Yeung, B.L. Yeo, and B. Liu, "Extracting Story Units from Long Programs for Video Browsing and Navigation," IEEE Multimedia 96 Proceedings, pp.296-305, 1996.
- [3] H. Aoki, S. Shimotsuji, and O. Hori, "A Shot Classification Method of Selecting Effective Key-Frames for Video Browsing," ACM Multimedia 96 Proceedings, pp.1-10, 1996
- [4] 青木恒, 下辻成佳, 堀修「映像ブラウジングのための類似ショット統合」情処研究報告 96-HI-67, 1996.
- [5] 金子敏充, 堀修「動きベクトル符号量を用いたMP EG映像からの高速カット検出」信学会研究報告 PRMU96-100, 1996.
- [6] M.M. Yeung, B.L. Yeo, W. Wolf, and B. Liu, "Video Browsing using Clustering and Scene Transitions on Compressed Sequences," SPIE Intl. Conference on Multimedia Computing and Networking Proceedings, vol.2417, pp.399-413, 1995.
- [7] ISO/IEC 13818-2: "Information Technology — Generic Coding of Moving Pictures and Associated Audio: Video," International Standard, 1995.