固定長の時空間画像に基づく映像クラスタリング

6L−1

岡本 啓嗣　八杉 将伸　馬場口 登　北橋 忠宏
大阪大学 産業科学研究所

1 はじめに

ＴＶなどの映像メディアを有効に，効率良く利用するためには，映像を部分映像に分割し，その映像の内容に基づいて構造化する必要がある．映像を構造化することができれば，所望の映像に対して効率良くアクセスすることができる．そこで，本研究では構造化における部分映像間の関係の一つである類似関係に着目し，映像中の類似部分をクラスタリングする手法の開発を試みる．従来，映像クラスタリングではショットを単位として，類似したものを統合していくことが多いが [1][2]，その境界の完全な検出技術が確立されていない．そこで本手法では，映像を一定時間毎に分割した固定長の映像セグメントを単位とし，時間軸方向への特徴を陽に表現するため，時空間画像を生成し，その同時生起行列を用いて映像セグメントをクラスタリングする．

2 手法概略

本手法の概略を図 1 に示す．まず映像を一定時間で分割し，そのときの固定長の映像を映像セグメントと呼び，これを単位とする．次に各映像セグメントから時空間画像を生成し，その時空間画像から特徴ベクトルとして HVC の値による同時生起行列を取り出す．そしてその特徴ベクトルを用いて次のような手順でクラスタリングする．まず，映像セグメントに複数のショットが含まれているかどうかを判定する．そのような映像セグメントを「非均質である」と定義し，非均質なセグメントを集めて一つのクラスタを形成する．次に残りの均質なセグメントを特徴ベクトル間の距離を用いてクラスタリングする．

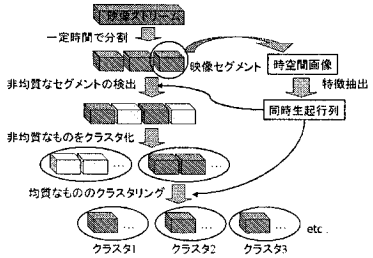

図 1: 手法概略図

3 時空間画像からの特徴抽出

3.1 時空間画像の作成

画像ストリームはフレームが時間軸方向に連続する時空間である．よって，フレーム画像に x 軸，y 軸を設け，時間方向に t 軸を設けると，画像ストリーム中の座標 F は $F=(x,y,t)$ で表される．ここでは，図 2 のような 4 つの平面を考える．すなわち，

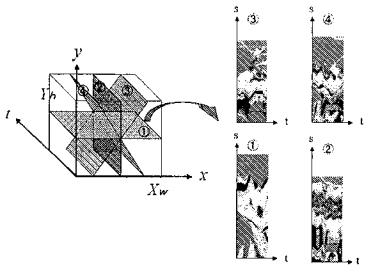

図 2: 時空間画像

① $T_h(s,t)=F(X_w-s,Y_h/2,t)$
$\{0\le s\le X_w,\ 0\le t\le T\}$

② $T_v(s,t)=F(X_w/2,s,t)$
$\{0\le s\le Y_h,\ 0\le t\le T\}$

③ $T_{d1}(s,t)=F(s,s-X_w/2+Y_h/2,t)$
$\{0\le s\le Y_h,\ 0\le t\le T\}$

④ $T_{d2}(s,t)=F(s,-s+X_w/2+Y_h/2,t)$
$\{0\le s\le Y_h,\ 0\le t\le T\}$

ここで，X_w はフレームの幅，Y_h はフレームの高さ，T は映像セグメント中のフレーム数である．このような時間軸に平行な画像ストリームの断面に現れる画像を時空間画像と呼ぶ．

3.2 同時生起行列

画面の変化が大きい映像の場合，時空間画像上で時間軸方向に変化が生じる．逆に静止している映像ならばそのような変化は現れない．これをテクスチャ特徴として取り出すため，同時生起行列を用いる．同時生起行列とは，画像の濃度 i の点 p_1 から一定の距離 r，角度 θ の相対位置 $\delta=(r,\theta)$ に濃度 j の点 p_2 $(p_1,p_2\in T(s,t))$ が出現する頻度を $P_\delta(i,j)$ とし，一定の δ について，全ての濃度の組合せ (i,j) に対して $P_\delta(i,j)$ を求めて行列 A で表したものである．濃度の階調数を n とすると，A の要素 a_{ij} は次のように書ける．

$$a_{ij}=P_\delta(i,j)\ \ \{0\le i\le n,0\le j\le n\}$$

3.3 特徴抽出

4 つの時空間画像に対して，前項に挙げた同時生起行列を，RGB 色空間における各要素の濃度からではなく修正マンセル色空間の要素である H(色相)V(明度)C(彩度) を 10 階調 $(n=10)$ に正規化したものから生成する．これは，RGB 色空間では各要素どうしの相関が高いことから，各々について同時生起行列を生成してもそれほど違いが生じないためである．そして画像ストリームにおいて次のフレームの同じ位置にあ

Video Clustering Based on Spatio-Temporal Image with Fixed Length
Hirotsugu OKAMOTO, Yukinobu YASUGI,
Noboru BABAGUCHI and Tadahiro KITAHASHI,
ISIR, Osaka University

る画素との関係を表すために，相対位置を $\delta = (1,0)$ として同時生起行列を生成する．これにより，例えば画像の動きが小さいセグメントの時空間画像では，隣接したフレームの画素値が大きく変化しないために，対角線上に要素の集中した同時生起行列が生成され，動きの大きなセグメントの時空間画像では画素値に頻繁に変化が生じるために，要素の分散した同時生起行列が生成される．このように，同時生起行列により色の頻度情報と動きに関する情報の両方を得ることができる．

このようにして各セグメント中の4枚の時空間画像から生成した同時生起行列を基に，画素数で正規化する．これにより，h_{ij}, v_{ij}, c_{ij} を各々H,V,C の同時生起行列の要素として次の式で表される特徴ベクトル $\mathbf{X}$ が得られる．

$$\mathbf{X} = (h_{00}, \ldots, h_{nn}, v_{00}, \ldots, v_{nn}, c_{00}, \ldots, c_{nn})$$

4 クラスタリング手法

4.1 非均質なセグメントの検出

非均質な映像セグメントは，内部にショットを複数含むため，他のセグメントとは性質が異なる．よってまず非均質なセグメントを一つのクラスタにまとめる．

映像セグメントが非均質であれば，複数のショットの一部がセグメント中に含まれており，そのショット境界の前後の時空間画像のパターンに差異が生じる．そこで，映像セグメントを時間軸方向に半分に分割して前後の時空間画像から同時生起行列を生成し，それぞれ特徴ベクトル $\mathbf{X}_1, \mathbf{X}_2$ とする．d をユークリッド距離として，$d(\mathbf{X}_1, \mathbf{X}_2) < \tau$ ならば非均質であるとする．しかし，等分しただけでは，中央付近にショット境界がある場合にしか対応できない．そこで，再帰的に等分を繰り返し，映像セグメントの端に境界が存在する場合にも対応する．このようにして検出された非均質なセグメントを一つのクラスタにまとめる．

4.2 均質なセグメントのクラスタリング

映像セグメント全体に対して取り出した時空間画像から同時生起行列を生成し，それを各映像セグメントの特徴ベクトルとしてクラスタリングする．クラスタ間の類似度を次のように定義する．

$$d_{max}(C_k, C_l) = \max_{\mathbf{X} \in C_k, \mathbf{Y} \in C_l} d(\mathbf{X}, \mathbf{Y})$$

$C_{k(l)}$ は $k(l)$ 番目のクラスタ，$\mathbf{X}, \mathbf{Y}$ はクラスタの要素である特徴ベクトル，$d(\mathbf{X}, \mathbf{Y})$ は要素 $\mathbf{X}$，$\mathbf{Y}$ 間のユークリッド距離である．

クラスタリングには，最も類似度の高いものから順に統合していく階層的クラスタリングアルゴリズムを用いる．クラスタの統合を止める閾値は，すべての映像セグメント間の平均類似度とした．これは，いくつかの映像に対して，すべての映像セグメント間の類似度を求め，その分布を調べた結果，中央付近にピークを一つもつ分布となったためである．

5 実験及び検証

次の TV 映像に対して実験を行った．映像 1〜3 はアメリカンフットボール，映像 4,5 は野球，映像 6,7 はニュース，映像 8,9 はバラエティである．これらの

表 1: 各映像の正解率

映像	均質	非均質	全体
1	75% (92/122)	79% (22/28)	76%
2	82% (94/114)	78% (28/36)	81%
3	71% (82/116)	68% (23/34)	70%
4	86% (97/113)	92% (34/37)	87%
5	74% (63/85)	88% (57/65)	80%
6	83% (103/124)	85% (17/20)	83%
7	76% (41/54)	90% (18/20)	80%
8	73% (72/98)	69% (36/52)	72%
9	77% (89/115)	63% (22/35)	79%
全体	78% (733/941)	79% (257/327)	78%

フレームレートはすべて 30 フレーム/秒 であり，フレームサイズは 縦 × 横 =240×320 である．また，映像中によく出現する短いショットの長さが 4 秒程度であったため，映像セグメントの時間単位を，その半分である長さの 2 秒に設定した．

セグメントが非均質なクラスタに属するかどうかの判定に用いた閾値は，$\tau = 0.35$ である．これは，数本の映像データからランダムにピックアップした部分に対して正解率が最も良かったものを適用した．なお，正解率とは次の値である．

$$\text{正解率} = \frac{\text{クラスタに正しく属するセグメント数}}{\text{全セグメント数}}$$

結果を表 1 に示す．野球，ニュースといった場面ごとの差異の大きな映像に関しては良好な結果を得た．精度を下げた要因に，アメリカンフットボールにおける，選手が多く写る中距離の映像と，アップ映像の区別がある．これは，画面の大きな領域に対して時空間画像を取ることにより，背景が多く写っている場合に特徴が類似してしまうためである．これに対応する方法として，一つは画面中央の注視領域に限って時空間画像を取り，背景の悪影響を防ぐことが考えられる．また，人物が多く写る中距離の映像のほうがフレーム画像の複雑さが大きいことから，時空間画像の $\delta = (1,90)$ の方向にも同時生起行列を生成し，時間軸方向だけでなく，フレーム画像内での複雑さも特徴とすることにより，アップ映像との差異を大きくする方法も考えられる．

6 むすび

本稿では，ショットという可変長の映像単位とは異なり，固定長の映像セグメントを単位として，一定時間毎に映像セグメントに分割し，それをクラスタリングする手法を提案した．その映像セグメントの特徴として，時間軸方向への特徴を陽に表現するため，時空間画像を生成し，その同時生起行列を使用する手法について述べた．本手法を様々な映像に適用したところ，正解率 78%とほぼ良好な結果が得られた．

尚，本研究の一部は，日本学術振興会科学研究費・基盤 (B)(代表:馬場口) の補助による．

参考文献

[1] M.M.Yeung and B.L.Yeo ,"Time-constrained clustering for segmentation of video into story units," in *Int.Conf.PatternRecognition(ICPR '96)*, vol.C, pp.375-380, Aug.1996.

[2] 青木恒, 堀修,"繰り返しショットの統合による階層化アイコンを用いたビデオ・インタフェース," 情報処理学会論文誌, Vol.39, No.5, pp.1317-1324, May.1998.