I-008

階層的クラスタリングを利用した映像ショット検出の一検討
A Study on Video Shot Boundary Detection by Hierarchical Clustering

梅田 直樹† 青木 輝勝‡ 沼澤潤二‡
Naoki Umeda Terumasa Aoki Junji Numazawa

1. はじめに

近年，データの圧縮技術やネットワーク関連技術，記憶媒体の研究成果により，個人が扱える映像コンテンツは膨大なものとなっている．そのため，膨大な数の映像コンテンツの中から，視聴や再利用のために目的の映像コンテンツのシーンやショットを探すことは非常に困難となっている．そのため，映像コンテンツに対して，あらかじめ索引情報等のメタデータをつけることで，意味内容に基づく検索を行うための研究が盛んに行われている．

メタデータを付与する前処理として，映像の構造化が必須であるといわれている．本研究ではその構造レベル(フレーム，ショット，シーン，クリップ)の中の，ショットレベルの構造化を行うためにショットを検出することを目的としている．

筆者らは特に検出漏れをゼロに保ったまま，誤検出を減らすことを目指した手法の研究を進めている．そのため，フレーム単位で種々の画像特徴量を抽出しその類似度によりフレームのクラスタリングを行い，ショット検出する手法を提案してきた．ここでは検出精度を上げるために，映像フレームの集合は連続データであるということを考慮してクラスタリングを行いショット境界検出を行った結果を報告する．

2. 従来研究の問題点

2007 年の国際的な動画検索を対象とするワークショップである TRECVID[1]で行われたショット境界検出の成果を見ると，CUT に対しては Recall，Precision 共に 98%近い結果が発表されている．また，GT (Gradual Transition)では，最も良い結果でも Recall，Precision 共に 80%を超えるものはない．

検出漏れと誤検出の原因として従来研究でいくつか指摘されてきた．しかし，筆者らが従来研究で最も問題であると考えたことは，検出率を上げる過程で，検出漏れが存在していることである．

メタデータ付与のために，ショット境界検出で処理された結果を用いるときに，検出漏れが 1 箇所でもある場合その漏れを探すコストは，誤検出を探すコストに比べて非常に高いと考えられる．なぜなら検出漏れがある場合は映像コンテンツを再び最初から最後まで見る必要があるためであり，誤検出がある場合の検出されたショット境界からその誤りを探すコストに比べると検出漏れがある場合のほうがより問題となると考えるからである．

†東北大学大学院情報科学研究科 Tohoku Univ. Graduate School of Information Sciences
‡東北大学電気通信研究所 Tohoku Univ. RIEC

3. 提案手法

本研究では，映像クリップからフレーム間差分の数値を分類してショット境界を分割して行く従来研究で主に使われている手法を採らない．全てのショットが，一枚のフレームという状態，つまり検出漏れがゼロの状態から同じショットであるフレーム同士を繋げていくことで，検出漏れがゼロの状態を保ったまま，ショット自体を検出する手法を提案する．

また，検出の精度を上げるためにショット境界検出に特化したベクトル空間を見つけることを目指す．そのために，フレームから抽出できる画像特徴量をできるだけ多くショット境界検出のシステムに組み込み，次元縮約を行うことを提案する．次元縮約とは，本質的な情報を保持したままより低次元のベクトル空間に変換する処理のことである．これにより，低次元のショット境界検出に特化したベクトル空間を発見することができると考えられる．

以上の提案を盛り込んだシステムを図 1 に示す．また，以下でブロックの詳細について説明する．

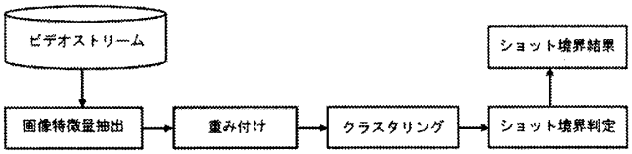

図1 提案システムの概要

3.1 画像特徴量抽出

検出漏れを避けるためにフレーム同士の差を十分に表すことができるように多くの画像特徴量を使うことにする．そのために利用する画像特徴量は RGB カラーヒストグラム，HSV カラーレイアウト，エッジヒストグラムである．

RGB カラーヒストグラムではフレームを 4×4 に分割し，それぞれのブロックに対して各ピクセルを各ビンに振り分ける．HSV カラーレイアウトでは 16×16pixels のブロックごとに平均色を計算する．また，エッジヒストグラムの抽出手法として，Park らによる手法[2]を用いた．

加えて，クラスタリングにより分類する対象は映像のフレーム毎の特徴量ベクトルである．そのため，フレームは連続データであることを考慮してタイムスタンプを特徴量として用いる．

3.2 重み付け

どのような画像特徴量を利用するかと同様に，どのような距離行列を利用するのか，特徴量毎にどのような重

み付けを行うかといったことは結果に多大な影響を与える．

ここでは，事前にショット毎のクラスタに分けられた教師データを基に画像特徴量毎の重み付けを決定する．この手法はウォード法のクラスタ間の非類似度を利用した手法である．ショット毎に分けられたクラスタ C_i , C_j に対して，次式で[0, 1]に正規化された変数 k の重み付け係数 w^k を計算する．

$$w^k(C_i, C_j) = \frac{E^k(C_i \cap C_j) - E^k(C_i) - E^k(C_j)}{|C_i \cap C_j|}$$

$E^k(C)$ はクラスタ C における変数 k の平均と各個体との差の 2 乗和を表しており，$w^k(C_i , C_j)$ の値が高いほど変数 k においてクラスタ C_i , C_j の非類似度が高いといえる．

クラスタ C_i , C_j の組み合わせは多数あるが，ここでは隣接したショットのクラスタについて計算した $w^k(C_i , C_j)$ の平均を変数 k の重み付け係数 w^k とする．

3.3 クラスタリング

フレーム毎の特徴量ベクトルを用いて，類似フレーム同士を併合していき，クラスタと呼ばれる集合を作る．このような処理をすることで，同じショット内のフレームは同じクラスタに分けられ，ショット自体を検出することができる．

ここではクラスタリングを行う対象が連続データであることに着目した制約を加えた制約付きクラスタリングを用いることを考える．具体的には，2 つのクラスタの中に連続したフレームがあるクラスタ対のうち，非類似度が最も低いクラスタから順次併合していくといった制約を設ける．このように併合するクラスタに制約を設けることにより，時系列的に遠いフレームが早い段階で併合されることを防ぐことができると考えられる．

クラスタ間の非類似度は凝集型階層的クラスタリングの手法であるウォード法を利用する．距離行列の再計算は Lance-Williams[3]の手法を用いる．クラスタリングを行う際，再生時間の長い映像クリップに対して一度にクラスタリングを行うと計算時間が長くなるため，適当な長さに分割してクラスタリングを行う．

また，非類似度の増加分が閾値を超えた箇所でクラスタを分割する．

3.4 ショット境界判定

クラスタリングによって、クラスタ毎に分けられたフレーム群を，フレーム毎のタイムスタンプ等を考慮してショットに分割する．ショット境界は異なるショットの境目なので，ショット毎に分割されたフレームからショット境界を検出することができる．

4. 提案手法の評価

提案手法の評価を行うために，評価実験を行った．ショット境界検出の評価尺度として，Precision と Recall が用いられている．それぞれの計算方法は次式で与えられる．

$$precision = \frac{D}{D + D_F}, recall = \frac{D}{D + D_M}$$

D は正しく検出されたショット境界の数，D_F は誤検出の数，D_M は検出漏れの数である．一般に，Precision と Recall は Trade-off な関係である．著者らは Recall を 1，つまり検出漏れが 0 の状態を保ちつつ，Precision を上げることを目指している．

3.1 節で説明した画像特徴量を用いて，3.1 節で説明した特徴量としてタイムスタンプを利用したクラスタリング手法，3.2 節で説明した重み付け，3.3 節で説明した連続データを考慮した制約付きクラスタリング手法の評価を行った．また，重み付けを行わなかった場合はそれぞれの画像特徴量の重みが等しくなるように正規化を行った．重み付けの評価実験では，映像クリップの半分を重み付けのための学習データとして交差検定を行った．

映像ソースとして 29 本の映像クリップ(総フレーム数 77,078 フレーム,CUT 数 328 ヶ所,GT 数 34 ヶ所)を用い，Recall=1 としたときの最大となる Precision を求めた．結果を表 1 に示す．

表 1　Recall = 1 としたときの最大 Precision

	重み付けあり	重み付けなし
タイムスタンプ	**0.418**	**0.363**
制約付きクラスタリング	**0.375**	**0.394**
タイムスタンプや制約付きクラスタリングを用いない場合		**0.332**

表 1 より重み付けを行い，タイムスタンプを利用したときが最も良い結果となった．また，表にはないが，タイムスタンプと制約付きクラスタリング，重み付けを行った結果は 0.373 となりあまり良い結果ではなかった．しかし，何れの結果でもタイムスタンプや制約付きクラスタリング，重み付けを用いない手法より高い Precision を得ることができた．タイムスタンプと制約付きクラスタリングはどちらとも時系列が近いクラスタを併合しやすくし，類似度が高くても時系列が遠いフレーム同士をつなげ難くするために導入した．そのため Precision が上がる結果につながったと考えられる．

5. まとめ

検出漏れをゼロに保ったまま誤検出を減らすことを目指した手法として，凝集型階層的クラスタリングによりショット内のフレームをつなぎ合わせ，ショット自体を検出することを提案してきた．評価実験から，HSV カラーレイアウトが用いる画像特徴量として最も適していることが分かった．

今後は，画像特徴量を増やし適切な特徴量を検討しより良い距離行列の計算手法を検討すると共に，ショット検出に特化したクラスタリング手法の提案を行い Precision の更なる改善を目指す．

参考文献

[1]TRECVID
http://www-nlpir.nist.gov/project/trecvid
[2]D. K. Park, etc."Efficient Use of Local Edge Histogram Descriptor", Proceedings of ACM International Workshop, (2000).
[3]G. Lance and W. Williams, "A general theory of classification sortingstrategies: 1. Hierarchical systems," *Comput. J.*, vol. 9, (1967).