階層的クラスタリングを利用した高精度ショット境界検出の一検討

梅田　直樹[1)]　青木　輝政[2)]　沼澤　潤二[3)]

[1)]東北大学　情報科学研究科　[2)3)]東北大学　電気通信研究所

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，情報ストレージ技術の進歩により，映像コンテンツの数は膨大なものとなっている．その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，意味内容に基づく映像コンテンツ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の映像ショット境界検出を目的とする．

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を研究している．そのため，フレーム毎のベクトルを用いた階層的クラスタリングを行う手法[1]を提案してきた．本論文では，クラスタリングを行うベクトルが映像の各フレームであるため，時間軸情報も考慮した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年に行われた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[2]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 ともに 0.98 に近い結果が報告されているが，GT(Gradual Transition)に関しては，最も良い結果でも Recall，Precision 共に 0.80 を超える結果は報告されていない．

これらの結果について，筆者らが最も問題であると考えたことは，検出率を上げる過程で，検出漏れを許容していることである．映像コンテンツのすべてのショットにメタデータを付与するには，ショット境界検出で処理された結果に検出漏れが1箇所でもある場合その漏れを人間の手で探すコストは，一般に誤検出を探すコストに比べて非常に高くなってしまうからである．

3. 提案手法

映像クリップからフレーム間差分をショット境界とショット内に分類してショット境界を検出してい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している．

同じショットの時間軸上で近いフレーム同士は類似度が高いという性質がある．そのため，分類対象がフレーム毎の多次元ベクトルである階層的クラスタリングを行うことにより，類似度が高いフレーム同士を繋げる処理を行う．しかし，クラスタリングを行う対象が映像の各フレームであるということ，つまり時間軸に対しての情報を考慮していないため，次のような問題がある．

- 時間的に不連続なフレームは同一ショットになるはずがないが，類似度が高いことで同一ショットになってしまう．
- 時間軸上で隣接し，類似度が低い2フレームが同一ショットになるはずはないが，これらのフレームがそれぞれ含まれる2クラスタ間の類似度が近くなるために，同一ショットに判定されてしまうことがある．

それぞれの問題を解決するために 3.1 と 3.2 で説明する手続きを加える．

3.1. 併合するクラスタの制約

階層的クラスタリングではクラスタ間距離が最も低い2つのクラスタから随時併合していっている．しかし，時系列的に遠い位置にあるフレームでも画像の類似度が高い場合は併合してし

"A Study on High Quality Video Shot Boundary Detection Method by Aggregative Hierarchical Clustering"

1)Naoki UMEDA・Tohoku Univ. Graduate School of Information Sciences

2)Terumasa AOKI・Tohoku Univ. RIEC

3)Junji NUMAZAWA・Tohoku Univ. RIEC

まうと考えられる（図 1）．その場合，クラスタの中のフレームが全て同一ショット内のフレームのまとまりとならない場合ができてしまう．

そのため，タイムスタンプが近いクラスタ群同士から併合していく必要がある．そこで，2 つのクラスタ内のフレームの内最もタイムスタンプが近いフレームのタイムスタンプの差が n フレーム以下であるクラスタ対を併合するクラスタ対の候補とし，候補以外は併合しない．

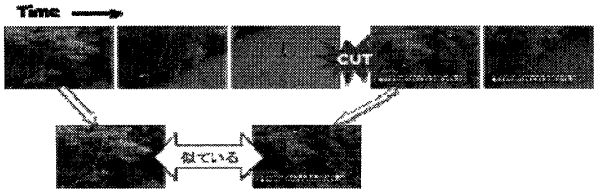

図 1　不連続なフレームを併合する問題

3.2. クラスタ間距離 D'の定義

図 2 のように隣接するフレームの類似度が低いにも関わらず，それぞれのフレームが含まれるクラスタの類似度が高くなってしまうために，同一ショットと判定されてしまう問題がある．

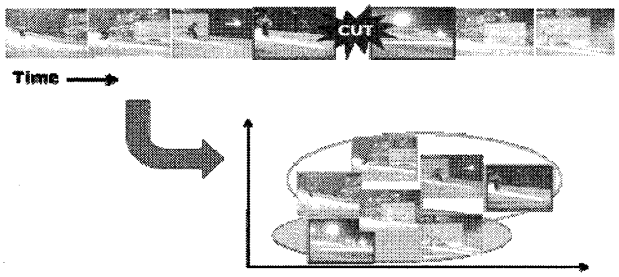

図 2　類似度の低い隣接するフレームを併合する問題

この問題を解決するために，クラスタ C_A, C_B に対して既存のウォード法のクラスタ間距離 $D(C_A, C_B)$に加えて，近接フレーム間の距離 $FD(C_A, C_B)$を導入して，α を用いた新しいクラスタ間距離 $D'(C_A, C_B)$を提案する．

$$FD(C_A, C_B) = \frac{|C_A \cup C_B|}{|F_A \cup F_B|} D(F_A, F_B)$$

$$D'(C_A, C_B) = \alpha D(C_A, C_B) + (1-\alpha) FD(C_A, C_B)$$

但し，$0 \le \alpha \le 1$

$|C_A \cup C_B|$は C_A, C_B に含まれるフレームの数である．また，F_A, F_B は C_A, C_B のクラスタ間でタイムスタンプが近い方から m フレームを C_A, C_B それぞれから取り出したものである．

4. 評価実験

併合するクラスタの制約とクラスタ間距離 D' に関して，その有効性を確かめる．

画像特徴量として MPEG-7 で標準化されているスケーラブルカラー，カラーレイアウト，エッジヒストグラムを用いた．　映像ソースとして 7 本の映像クリップ(総フレーム数 21,995, CUT 数 77, GT 数 21)を用いた．

クラスタ間距離が閾値を超えたところでクラスタに分割する．閾値は手動により検出漏れがゼロで Precision が最大となる値にした．

制約するフレーム数を 3 フレームとしたとき，Precision の値は制約を加えなかったときの 0.249 から 0.322 となり，制約を加えることの有効性が確かめられた．

また，クラスタ間距離 D'の近接フレーム数 m と α を変化させたときの結果を図 3 に示す．但し，制約するフレーム数は 3 フレームとした．

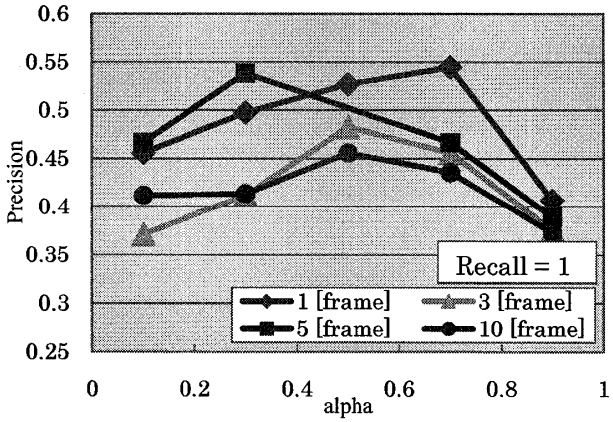

図 3　新しいクラスタ間距離 D'を用いたときの Precision

図 3 より，クラスタ間距離 D'を用いた時，Precision は最大 0.222 上がり，0.544 という結果が得られた．よって，クラスタ間距離 D'を用いることの有効性が確かめられた．

5. まとめ

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的とした階層的クラスタリングを行う手法を提案してきたが，既存の階層的クラスタリングに時間軸情報も考慮した手続きを加えることにより，ショット検出に有効なクラスタリングが行えることを確認した．

参考文献

[1]梅田直樹, 青木輝勝, 沼澤潤二, "フレームベースクラスタリングを利用したショット検出手法の一検討", 情報処理学会全国大会講演論文集, May 2009

[2]TRECVID
http://www-nlpir.nist.gov/project/trecvid