Vol.2009-AVM-67 No.9
2009/12/7

フレームベースクラスタリングを利用した高精度映像ショット検出の一検討

梅田直樹[†]　青木輝勝[††]　沼澤潤二[††]

映像の基本処理の 1 つにショット境界検出があるが，実用を考えた場合既存研究ではその精度が十分ではない．筆者らは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，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について検討を進めている．
本稿では筆者らの提案手法を基礎としてさらに検出精度を上げるために，クラスタリングを行う対象が映像の各フレームであることとし、これにクラスタリング技法を適用することによりを行い，これにより高精度なショット境界検出を行う手法についての実験結果を報告する．

A Study on High Quality Video Shot Boundary Detection Method by Frame-based Clustering

Naoki Umeda[†]　Terumasa Aoki[††]　and Junji Numazawa[††]

One of the fundamental image processing is shot boundary detection, but existing methods are not enough about their performances from practical points of view. We are researching a new shot boundary detection method based on frame-based clustering, which aims at decreasing false detection while keeping zero leakage (no overlooking of shot detection).
In this paper, we propose a new clustering method in order to improve the performance, and report some experimental results about this method. In this method, each frame in video streams is regarded as one element in a multi-dimensional clustering space and this method detects shot boundaries accurately from this multi-dimensional clustering space by using a good clustering algorithm.

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，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，クラスタリングを行う対象が映像の各フレームであることを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．

本論文の構成は以下のとおりである．2 章ではショット境界検出へのこれまでの取り組みについて触れ，その問題点について述べる．3 章では本研究で提案するフレームベースクラスタリングの概要を述べ，4 章から 6 章で提案手法を用いたショット境界検出のシステムについて説明する．7 章で評価実験結果を報告し，8 章でまとめと今後の課題を述べる．

2. 従来研究の問題点

ショット境界は境界の長さによって CUT と GT（Gradual Transition）に分類される（図 1）．連続する 2 つのショット S_i，S_j において，S_i の最終フレームと S_j の開始フレームの差を境界の長さとし，境界の長さが 1 フレームのものを CUT，2 フレーム以上のものを GT と呼ぶ．CUT は最終フレームと開始フレームを単純につなげたものであるのに対し，GT は 2 つのショットをなんらかのエフェクトにより滑らかにつなげているため，境界内の連続するフレームの変化が CUT と比べて少ない．加えて，エフェクトの種類はディゾルブ，フェード，ワイプなど多岐にわたる．そのため，GT は既存のフレーム間の差分を用いた手法では検出しにくい．

[†] 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

Vol.2009-AVM-67 No.9
2009/12/7

図 1　CUT と Gradual Transition

2007 年に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 0.98 に近い結果が発表されている．また，GT では，最も良い結果でも Recall，Precision 共に 0.80 を超えるものはない．

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．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人間が見てショット境界検出漏れを発見す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検出を探す時間と困難さに比べると検出漏れがある場合のほうがはるかに困難でありより多くの時間がかかるからである．

3.　提案手法の概要

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．（図 2）

検出の精度を上げるためにフレームベースクラスタリングのブロックでは，クラスタリングを行う対象が映像の各フレームであることに着目したクラスタリングを行う．

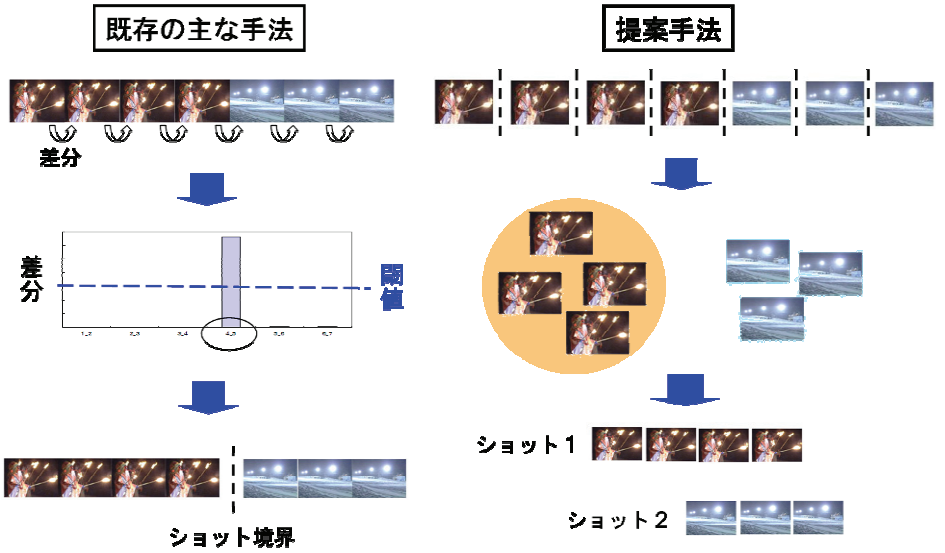

図 2 既存手法と提案手法

また，ショット境界判定のブロックでは，クラスタリングを行った結果として得られるショットからショット境界を見つける際に，フラッシュ等のノイズを排除する処理を行うことによって検出の精度を上げている．

以上の提案を盛り込んだシステムを図 3 に示す．また，4 章から 6 章で各ブロックの詳細について説明する．

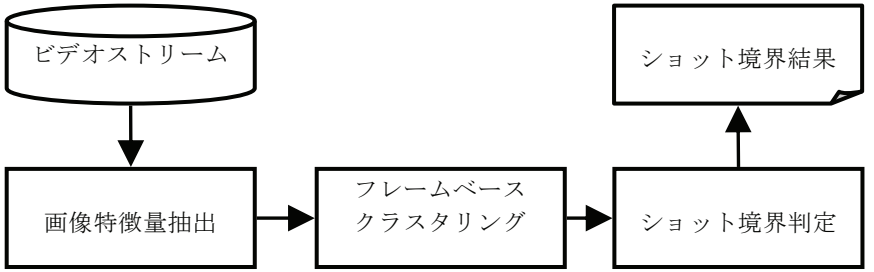

図 3　提案システムの構成

4. 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とする．そのために利用する画像特徴量は図 4 の HSV カラーヒストグラム，RGB カラーレイアウト，エッジヒストグラムの 3 種類の特徴量である．

既存の主な手法と異なるのは，ある 2 つのフレーム間の差分特徴量を利用することではなく，フレーム画像から抽出される上の画像特徴量をそのまま用いてクラスタリングを行うことにある．著者らはこの手法をフレームベースクラスタリングと呼ぶことにする．

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．また，用いたフレーム画像は 352×240pixels であり，それぞれの特徴量は正規化してある．

4.1 節から 4.3 節で各画像特徴量を説明する．

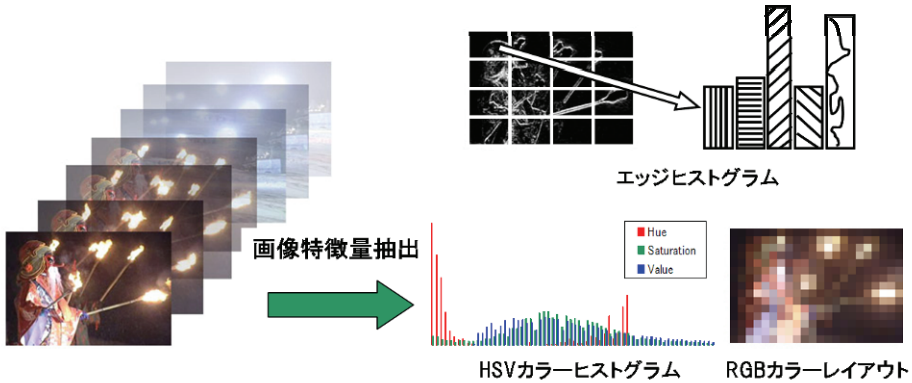

図 4　画像特徴量

4.1 HSV カラーヒストグラム

HSV カラーヒストグラムはショット境界検出の既存の手法である 2 つのフレームの差分を用いた手法で利用されることが多く，フレーム同士の類似度をよく表現できる．本研究では色空間として HSV 色空間を採用し，色相（45 段階），彩度（64 段階），明度（64 段階）の各ビンに対して各ピクセルを振り分ける．H と S，V の段階が異なるのは，画像処理に用いている OpenCV ライブラリ[2]に合わせて段階の数を決定しているからである．HSV ヒストグラムは 45 + 64 + 64 = 173 次元の画像特徴量となる．

4.2 RGB カラーレイアウト

RGB カラーレイアウトでは 16×16pixels のブロックごとに計算する．色空間として RGB 色空間を採用し，各ブロックの RGB それぞれに対して平均の値を計算する．よって，RGB カラーレイアウトは 22×15×3 = 990 次元の画像特徴量となる．

4.3 エッジヒストグラム

エッジヒストグラムはマルチメディア用メタデータ表記方法の国際標準化規格である MPEG-7 Part3 visual で記述されるエッジヒストグラムとほとんど同じものである．異なる部分は，MPEG-7 の方では 8 レヴェルに量子化され記述されるが，本研究で用いる場合，それでは粗すぎるため量子化テーブルを利用して作成した関数を用いて出力された値を用いることとした部分である．

エッジ特徴量としてエッジヒストグラムを用いたのは，ショット境界検出で用いた特徴量を MPEG-7 のメタデータとして 2 次利用することを考えたからである．

エッジヒストグラム特徴は，画像を 4×4 の格子状にブロック分割して，その各ブロックから，垂直エッジ，水平エッジ，角度 45° のエッジ，角度 135° のエッジ，方向性のないエッジの 5 種類のエッジの量を特徴量とする．つまり，エッジヒストグラムの特徴量は 4×4 それぞれのブロックに対して角度 5 種類の割合を出すために 80 次元の画像特徴量となる．

エッジヒストグラムの抽出手法として，Park らによる手法[3]を用いる．

5. フレームベースクラスタリング

フレーム毎の特徴量ベクトルを用いて，1 つのフレームが 1 つのクラスタとなっている状態からクラスタ間の距離を計算し，その距離が小さい 2 つのクラスタから随時併合していく．そのようにして，クラスタの数が 1 つになるまでクラスタ間距離の再計算と併合を繰り返す．併合過程は図 5 のようなデンドログラムで表され，適当なところで切断することにより，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．

クラスタ間の距離は凝集型階層的ク

ラスタリングの手法であるウォード法を利用する．クラスタ G_i と G_j のウォード法を用いたクラスタ間距離⊿E(G_i, G_j)は，クラスタリング対象 x と重心 M(G)を用いて式（1）で与えられる．

$$E(G) = \sum_{x_i \in G} \|x_i - M(G)\|^2$$
$$\Delta E(G_i, G_j) = E(G_i \bigcup G_j) - E(G_i) - E(G_j) \quad (1)$$

Vol.2009-AVM-67 No.9
2009/12/7

距離行列の再計算は Lance-Williams[4]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．

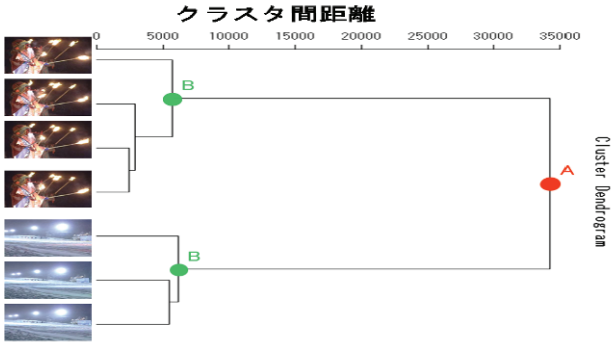

図 5　デンドログラム

5.1　併合するクラスタの制約

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．

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，このようにした場合，時系列的には遠い位置にあるフレームでも画像特徴量が似通っている場合は併合してしまうと考えられる．クラスタリングを行う際の特徴量ベクトルにタイムスタンプを用いることにより，タイムスタンプが遠いフレームは併合しにくくなっているが，併合する際に制約を設けることにより，確実にタイムスタンプが遠いフレームを併合しないようにする．

具体的には，2 つのクラスタの中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閾値以下のクラスタ対の中で，クラスタ間距離が最も低いクラスタ対を併合するように制約を設ける．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．

ただし，閾値を 1 フレームとしたときは連続するフレームを持つクラスタ同士しか併合できないこととなり，6.1 節で説明するフラッシュ等のノイズ対策による処理がうまく作用しない，つまりノイズに弱いシステムとなってしまうことが考えられる．そのため，閾値をある程度大きい値とすることで，フラッシュ等のノイズに対して頑強なシステムとする必要がある．

5.2　切断箇所

非階層的クラスタリングではなく，階層的クラスタリングを利用する利点のひとつとして，あらかじめクラスタ数を決めなくて良いことが上げられる．非階層的クラスタリングの k-means 法の最も単純なアルゴリズム[5]では，クラスタの数が予めわかっている必要があるが，階層的クラスタリングでは得られたデンドログラムのクラスタ間距離の遷移からクラスタ数を決めることができる．ショットの数を予め知ることは困難であるため，クラスタ形成過程よりクラスタが決めることが重要となると考えられる．

ここでは，2 つのクラスタを併合したときのクラスタ間距離（例：図 5 の A）とそれぞれのクラスタの以前の併合でのクラスタ間距離（例：図 5 の B）の差を計算し，その差がどちらともある閾値を超えたところを切断箇所としている．この閾値を変えることにより，検出漏れと誤検出のバランスを変えることができる．本研究では，検出漏れがゼロとなる値を閾値としている．

6.　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．

6.1　フラッシュ等のノイズ対策

カメラのフラッシュ等の影響により，同じショット内であっても 1 から 3 フレームといった短い時間の間，フラッシュ以前のフレームとフラッシュ以降のフレームと類似度の低いフレームが現れる場合がある．(図 6) この場合，提案手法のフレームベースベクトルクラスタリングを行った結果として，フラッシュのあるフレームだけ異なるクラスタになってしまい誤検出となってしまう．

これを避けるため，3 フレーム以下のクラスタであり，そのクラスタの前後のフレームが所属するクラスタが同じであった場合に，3 フレーム以下のクラスタはノイズであるする．そして，前後と同じクラスタ，つまりは同じショットとすることにより，フラッシュのようなノイズによる誤検出を排除する．

図 6　フラッシュライトによる誤検出例

Vol.2009-AVM-67 No.9
2009/12/7

7. 評価実験

7.1 評価方法

提案手法の評価を行うために，評価実験を行った．一般にショット境界検出の評価尺度として，Precision と Recall が用いられているため，本稿でもこれらを評価指標として用いることとする．それぞれの計算方法は式(2)で与えられる．

$$precision = \frac{D}{D + D_F}, recall = \frac{D}{D + D_M} \quad (2)$$

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 はトレードオフな関係である．著者らは Recall を 1，つまり検出漏れが 0 の状態を保ちつつ，誤検出を減らし Precision を上げることを目指している．

7.2 実験条件

評価データとして映像ソース 26 本を用いた．総フレーム数は 77,078，ショット境界数は，CUT が 328 ヶ所，GT が 33 ヶ所の計 361 ヶ所ある．ここでは，検出漏れがゼロを保った状態で最も誤検出が少ないように，切断箇所の閾値を設定したときの Precision の値を出す．

また，計算時間削減のために映像ソースを 180 フレームごとに分割してクラスタリングを行う．

7.3 実験結果

併合するクラスタの制約を 1，2，3，10，30，60，90 フレームと変化させ，それぞれのフレームでフラッシュ等のノイズ対策を行った場合と行わなかった場合の Recall=1 としたときの Precision の結果を表 1，図 7 に示す．

表 1 併合するクラスタの制約とノイズ対策を行ったときの Precision の値(Recall = 1)

併合するクラスタの制約[フレーム]		1	2	3	10	30	60	90	制約なし
ノイズ対策	○	0.401	0.405	0.403	0.383	0.365	0.352	0.345	0.344
	×	0.401	0.391	0.367	0.344	0.316	0.304	0.296	0.295

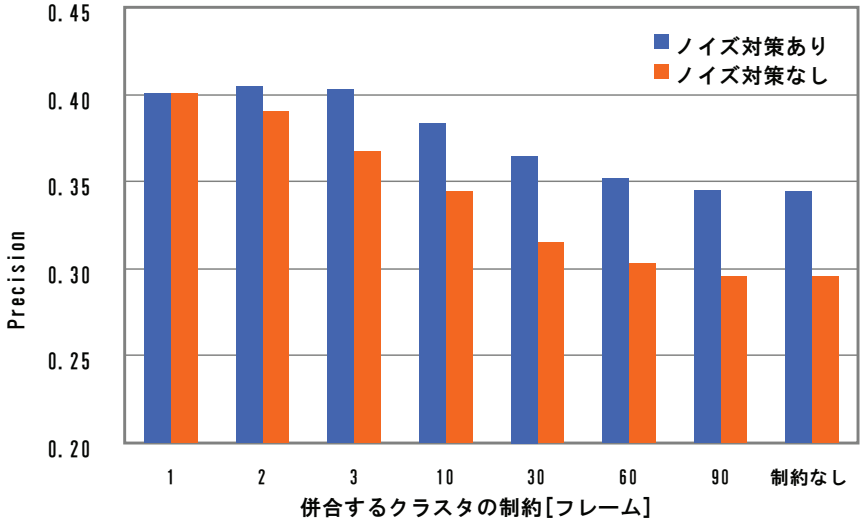

図 7 併合するクラスタの制約とノイズ対策を行ったときの Precision の値(Recall=1)

図 7 より，5.1 節の併合する際に制約を設けることはショット境界検出に対して，有効性が確認できる．これは，併合する順番を考えたとき近くのフレーム同士を併合しやすくする，つまり映像のショットということを考慮してクラスタリングを行うことが重要となることを示していると考えられる．

また，6.1 節のフラッシュ等のノイズ対策は併合する際の制約が 1 フレーム以上のときに誤検出を減らしていることがわかる．併合するクラスタの制約とノイズ対策を行うことにより，Recall=1 としたときの Precision の値が最大 0.11 上がり 0.405 となった．

次に提案手法と TRECVID2007 参加機関の CUT と GT の結果を合わせた ALL の結果と比較したものを図 8 に示す. 提案手法では表 1 で最も良い結果であった併合するクラスタの制約を 2 フレームとしノイズ対策を行った結果と，併合するクラスタの制約とノイズ対策を行わなかった結果をプロットした．

TRECVID2007 の結果では Recall = 1 を達成している結果はなかった．また，最も Recall が高い結果（Recall = 0.967）では Precision = 0.379 となり Recall = 1 を達成することが困難であることがわかる．

提案手法では Recall = 1 としたとき，Precision = 0．405 という結果となった．使っている映像ソース群が異なるため正確な比較はできないが，この Precision の値は TRECVID2007 の参加機関の結果と比較して低い値ではあるが，Recall=1 を実現したことは実用上極めて意義が高く，今後は Recall=1 を保ちながら Precision 値を高めることにより理想的な手法の提案が行えるものと期待できる．（図 8）

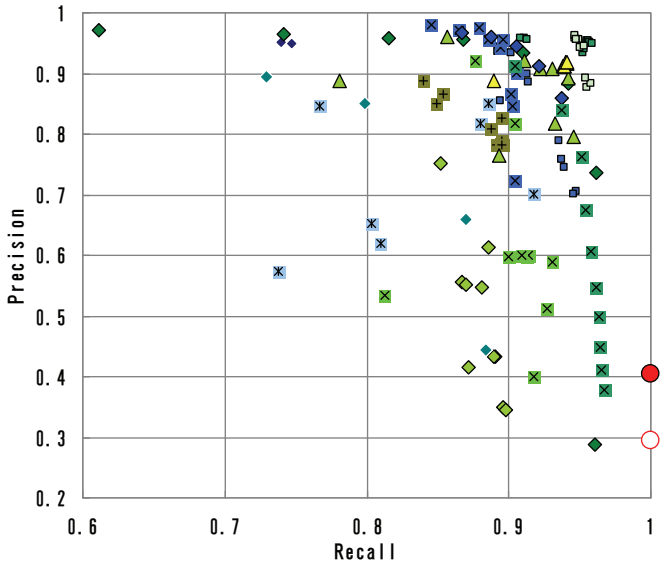

図 8　提案手法と TRECVID2007 参加者の ALL の認識結果
(●提案手法で最も良い結果，○が併合するクラスタの制約とノイズ対策を行わなかった結果，その他のマークが TRECVID2007 の各参加機関の結果)

8. まとめと今後の課題

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，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．今回はクラスタ同士を併合する際に，併合するクラスタ対が映像のフレームの集合であるということを考慮した制約を設けることで誤検出を減らし Precision をあげることができた．

今後の課題として次の 2 点を挙げる．まず，画像特徴量を増やし適切な特徴量を検討し次元縮約を行うことにより，Precision の更なる向上と計算時間の削減を目指す．提案手法では，すべての特徴量の重みを等しくなるようにしたが，ショット境界に必要な情報を保持したまま次元を減らすことにより，低次元のショット検出に適したベクトル空間が得られると考えられる．また，次元数が下がることで計算時間が削減できると期待される．

次に，本論で述べた実験結果から誤検出の部分を検出する後処理を導入することを検討したい．提案したシステムの場合，長いショットやカメラモーション，激しく動くオブジェクトがある映像では誤検出が多くなってしまう．そのため，そのような誤検出の原因となっている箇所を検出することにより Precision の更なる向上を目指すこととする．

参考文献

1) TRECVID2007
http://www-nlpir.nist.gov/project/trecvid
2) “Open Source Computer Vision”
http://opencv.willowgarage.com/wiki/
3) D. K. Park, etc., “Efficient Use of Local Edge Histogram Descriptor”, Proceedings of ACM International Workshop, (2000).
4) G. Lance and W. Williams, “A general theory of classification sorting strategies: 1. Hierarchical systems,” *Computer. J.*, vol. 9, (1967).
5) E.W.Forgy, “Cluster analysis of multivariate data: efficiency vs. interpretability of classifications”, biometrics, vol.21, pp. 768-769(Abstract) (1965).