2008－CVIM－163 (2)
2008／5／8

色，動き，顔特徴に基づくTRECVIDラッシュ映像の自動要約

野口顕嗣† 柳井啓司†

本研究では，国際映像処理ワークショップTRECVIDで2007年から始まった映像自動要約タスク(rushes summarization)について取り組む．映像中のショットを色，動き，顔特徴に基づいてクラスタリングし，代表ショットを選ぶことにより映像の自動要約を実現する方法について提案する．実験として最初に3つのシステムについて比較した．1つは特徴量が色だけのもの，2つめは特徴量として動きと色を用いたもの，最後に動き，色，顔を用いたものである．次にこれらのシステムとTRECVID 2007の参加者との結果を比較した．3つのシステムを比べた結果，動き情報を用いたものと用いなかったものでは結果に大きな差が表れた．次に顔情報であるが，これも結果に大きな差を与えた．以上のことから特徴に顔と，動きを加えることはこのタスクにおいてとても有効であることが分かった．ただしクラスタリングにおいては色特徴を使用しているので，全体的に色が変化しないビデオに関しては良い結果は出せなかった．また，ground truthとの一致率であるIN値に関してはTRECVID 2007の参加者と比べて良い結果が得られた一方，システムの実行時間は他の参加者と比べ良い結果を得ることができなかった．

Rushes summarization by color, motion and face

Akitsugu Noguchi† and Keiji yanai†

In this paper, we present a method for BBC rushes summarization which is one of a task of TRECVID. In the proposed method, first an input video is decomposed into shots by comparing consecutive frames. Then, these shots are grouped by the k-means method, using color feature, motion feature and face feature. In the experiments, we compared three systems which employed the following feature combinations: "color", "color and motion" and "color, motion and faces". Next we compared these results with ones of the participants of TRECVID 2007. As a result, we found that motion features and face features were effective. The inclusion rate with ground truth was relatively good, while the system time was not so good.

1. はじめに

1.1 背　景

インターネットの動画配信やデジタルビデオカメラの普及にしたがって一般ユーザも大量の映像を入手できるようになった．しかし一方で閲覧したい動画を探すことが大変困難な問題となってしまっている．また多くの映像がそのままの形でなく，なんらかの編集をされて配信されている場合が多くそれらの作業の多くは手作業で行われている．

重要な部分のみからなる映像を作っておくことによって，動画検索が比較的容易になる．さらに映像の編集はほとんどが手作業で行われているが，これからは更に映像は増えていくと考えられる．そうなれば，編集にかかる手間は膨大なものとなる．このことからも映像の自動要約に対するニーズが高まっていることが分かる．

1.2 目　的

本研究ではTRECVIDで2007年から始まったrushes summarizationを取り上げ，実際にTRECVIDで用いられたデータを利用し，ラッシュ映像を要約するシステムについて考える．そして実際のTRECVIDと同様の評価を行い，TRECVIDの参加者との結果を比較，考察する．

2. TRECVID

2.1 TRECVIDについて

TRECVIDとは，アメリカの国立標準技術研究所NIST(National Institute of Technology)の研究部門が行うテキスト検索ワークショップTREC(Text REtrieval Contest)から派生したビデオ映像検索ワーク

† 電気通信大学大学院 電気通信学研究科 情報工学専攻
Department of Computer Science, The University of Electro-Communications

ショップである．参加者は，TRECVID が配布する実験用データに基づいてタスクを取り組み，与えられた期限までに各タスクの実験結果を NIST に送付する．NIST では，タスクごとに設定された評価基準にしたがって，すべての実験結果を評価し，その結果を参加者に返送するとともに，TRECVID Workshop で報告を行う．

2.2 TRECVID 2007

2.2.1 タスク

TRECVID 2007 では，以下の 4 つのタスクが設定された．

- Shot boundary detection(ショット境界検出)
  実験対象映像ファイルに含まれる各ショット間の切り替え点を自動的に検出するタスク
- High-level feature extraction(高次元特徴抽出)
  指定された事象が出現する個所を，テスト画像データから検出するタスク
- Search(検索)
  与えられたクエリを満たすショットを効率的に検索するシステムを開発するタスク
- Rushes summarization(ラッシュ映像要約)
  未編集映像である，ラッシュ映像を決められた長さ以下に要約するタスク

本研究で今回取り組むタスクは rushes summarization である．

2.2.2 BBC Rushes Summarization

Rushes summarization は与えられたラッシュ映像 (MPEG-1) を決められた時間以下 (2007 においては 4%以下) に自動で要約するタスクである．ラッシュ映像とは，未編集の映像のことであり俳優の NG シーンなどの繰り返しのシーン，カメラが固定されていて長い間動きがないシーンを含んでいる映像のことである．

与えられたビデオデータは最小 3.3 分，最大で 34.6 分であった．また 50 のビデオがシステムの development data として，42 のビデオが test data として与えられていた．

このタスクにおける評価方法は，テキスト形式の ground truth の一致率，リッカート尺度による，要約としての見易さや無駄の少なさのような主観的なものと，システムの実行にかかった時間，審査官が審査にかかった時間，要約の長さなどの客観的ものがある[1]．

図 1 はラッシュ映像の 10 秒ごとのフレームと実際に要約したフレームの例，表 1 はこの動画に対応する ground truth の一部である．

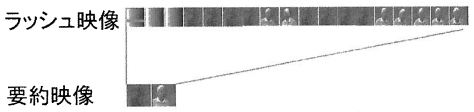

図 1 要約映像の 10 秒毎のフレーム

表 1 Ground truth の例

Shot of trees
Woman towards camera, stops and talks
Woman turns around and walks down footpath
Boy standing on bench facing yellow and pink puppets
Puppets standing behind bench, woman standing to left of bench facing puttets
Woman exits to left
Camera zooms in on two puppets
Yellow puppet exits to left
Red puppet picks up bad from bench and exits to left
Closeup of two puppets behind bench
Both puppets exit to left
Shot of footpath at side of road
Shot of road from different angle
Shot of tree-lined path
camera pans to right
woman and two puppets enter from right and exit on left
Shot of bush, with fence and tree in background
woman and two puppets enter from right, then exit to the left
Shot of yellow puppet beside bush, in front of house

3. 関連研究

TRECVID 2007 において 22 の参加者が様々な手法を提案している[1]．ここではそのなかで，Nii[2]，CMU[3]，hkpu[4] について紹介する．

Nii(国立情報学研究所)[2] では入力されたビデオを fragment という動きに重要な変化のない単位に分解し，それをクラスタリングを行い同一クラスタ内にある連続した fragment は同一 segment として併合される．そして隣接する segment 間の類似度を計算してこの値が閾値を下回っていた場合二つの segment を併合する．また新たな segment にたいして同一の作業を行っていく．最後に各クラスタにおいて最も長い segment を要約を製作するために選ぶといった手法を提案している．

他には，CMU(Carnegie Mellon University)[3] では，映像要約部と音声要約部のふたつからなるシステムを提案している．まず映像要約部であるがこれは最初にショット分割を行い，それぞれのショットの最初のフレームをキーフレームとして抽出し，それをもとに k-means でクラスタリングを行う．この時の k の値は要約ビデオの秒数となる．この時カラーバーなど

明らかに不必要なクラスが存在するのでそれらを除去する．その後再び k-means でクラスタリングを行い．それぞれのクラスタの重心に最も近いものを一秒選び要約とする．次に音声要約部であるが，編集した視覚特徴 ASR trsnscript の時間境界が一致した SNR 計算を用いて，スピーチが含まれているかを決定する．音声編集リストをそれぞれの視覚特徴の中間点に初期化した後，それぞれの音声をもっとも近い SNR 境界まで広げる作業を要約時間になるまで繰り返す手法を提案した．映像とのシンクロ性は失われるが要約理解の手助けとなる．

更に CMU では pan/zoom のカメラモーションを強調した手法を提案している．これはカメラモーションが検出された場合自動で 1 秒以上のタグをつける．そして上で述べたようにクラスタリングを行う．それぞれのクラスタにカメラモーションがあれば，最も長いものが代表と，そうでない場合顔が映っているものを優先的に，それも無い場合上で述べた方法でクラスタの代表を決定する．要約の時間が長かったらカメラモーションを要約したい時間になるまで短くする手法を提案した．

また hkpu(Hong Kong Polytechinc University)[4] ではこのタスクを 3 つのサブタスクに分けて考えている．1 つめはショット検出，2 つ目がジャンクショット検出と除去でカラーバーなどのショットは，色を用いて検出，クラッパーボードの検出のためにノックされた時に音響エネルギーが大幅に上がることを利用した新たな手法を提案している．3 つめが要約で方法として，まずそれぞれのショットを固定の長さに区切る．それぞれのクリップから最も静的であるものと最も動的であるものを選択する．この抽出されたフレームを時間順に並べ，そのなかで似ているフレームは除去される．残っているフレームから前後 12 フレームを探して要約に含める手法を提案した．

以上の研究を踏まえて，本研究では主に CMU の手法を参考にして，カメラモーションのみでなく動きの大きさに着目したショットのクラスタリングに基づく手法を提案する．

4. アルゴリズム概要

ここでは本システムのアルゴリズムの概要について説明する．図 2 はシステムの概要を表している．

CMU ではカメラモーションを人が認識するには 1 秒は短いとして，カメラモーションが含まれているシーンには 1 秒以上のタグをつける手法を提案しているが，本研究では，動作がひとつの動作は基本的に 1 秒以上かかると考え，前後するフレームの動き情報を計算し，一連の動作が終了するまで分割を待つという手法を提案する．

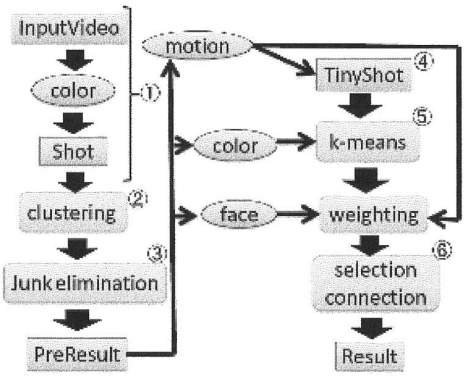

図 2 アルゴリズム概要

4.1 ショット分割 (図 2.1)

まず与えられたビデオを色特徴をもとに連続するフレームを比較していくことによってショット分割していく．またその際に各ショットにおける色特徴を計算する．

4.2 クラスタリング (図 2.2)

上で計算された色特徴をもとにクラスタリングを行う．この作業で似ているショットは同一クラスタになるので，繰り返し，ジャンクショットの除去が期待できる．

ただし各クラスタの代表は，最も長いものが選ばれる．

4.3 ジャンクショット検出 (図 2.3)

ラッシュ映像をみていくと図 3 に示すように明らかに不必要なフレームがあることが分かる．

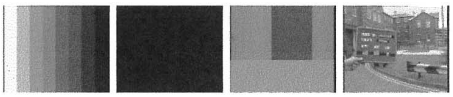

図 3 ジャンクショットの例

本システムでは右の 3 つのジャンクショットに関して色を用いて検出を行っている．クラスタの代表がジャンクショットの場合，そのクラスタはジャンクショットからなるクラスタとみなし除去する．

4.4 Tiny shot 分解 (図 2.4)

それぞれのクラスタの代表を秒単位に分解していく．またこの時に顔検出，色特徴の計算，動き情報の計算を行う．ただし 1 秒を過ぎていたとしても Lucas-Kanade 法[5] でオプティカルフローを計算してある程度の動きがあった場合は，一連の動作の途中であると

考え，動きが一定以下になるまで分解は行わないようにする．

4.5 K-means (図 2.5)

秒単位に分けられたビデオを色特徴を元に，オリジナルビデオの 4%以下になるように k の値を設定して，k-means アルゴリズムでクラスタリングしていく．

k-means アルゴリズムの手順は以下のようになる．

(1) K 個のクラスタの中心をランダムで決める
(2) それぞれのデータを最も近い中心と同一のクラスタへ割り当てる
(3) クラスタ毎に中心を計算し直す
すべてのクラスタ中心が変化しなければ終了
それ以外は 2 へ

4.6 Selection と connection (図 2.6)

各クラスタの代表は，できるだけ動きがあるもの，人が映っているもの が望ましいのでショットの中で顔が検出された場合そのショットには重み W($W > 1$) をつける．ただし実験においては W=1.5 である．また検出されなかった場合 W=1 である．この W の値とショットの動き情報を積算したものを重みとして，各クラスタから重みが最大のものを代表として選択し，これらを時間順につなぎあわせることにより要約映像とする．

5. 特 徴 量

5.1 グリッド分割したカラーヒストグラム

図 4 分割した画像

色特徴としては図 4 にしめすように 3 × 3 に分解した位置情報つき RGB カラーヒストグラムを使用する．そしてショットとしての色特徴は，式 1 で示す通りそのショットに含まれるカラーヒストグラムの平均であらわす．

$$C = \frac{1}{F}\sum_{i=2}^{F}\sum_{x=1}^{3}\sum_{y=1}^{3}\sum_{k=0}^{64} H_{ixy_k} + H_{i-1xy_k} \quad (1)$$

ただし F はショット中に含まれるフレームの数を，H_{ixy_k} はショット中の i 番目のフレームの格子 (x,y) のヒストグラムの k 番目の要素であることを示している．

ヒストグラム間の類似度はヒストグラムインターセクションで定義する．ヒストグラムインターセクションとは，図 5 に示す通り 2 つのヒストグラムの重なった部分を示す．式 2 で 2 つのヒストグラム間の類似度を定義する．

$$Sim_{i,j} = \sum_{x=1}^{3}\sum_{y=1}^{3}\sum_{k=1}^{64} min\{H_{ixy_k}, H_{jxy_k}\} \quad (2)$$

ショット分割の際にはこの Sim の値が閾値を越えた場合に分割を行う．

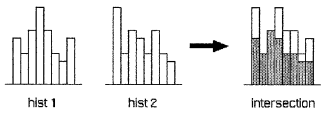

図 5 ヒストグラムインターセクション

5.2 動 き 情 報

5.2.1 オプティカルフロー

オプティカルフローとは視覚表現の中で物体の動きをベクトルで表したものである．抽出方法として主にブロックマッチング法や勾配法が用いられるが，本研究では勾配法の一種であり精度が良好でかつ処理が高速な Lucas-Kanade 法を用いて推定を行う．

5.2.2 Lucas-Kanade 法

勾配法には Horn & Schunck アルゴリズム[6] などがあるが，Lucas-Kanade 法を用いることで高速で精密なオプティカルフロー推定ができる．またこの方法は同一物体の局所領域内ではオプティカルフローはほぼ一様になると仮定する空間的局所最適化法の一つである．

画像座標と時刻 t の時 (x,y,t)，画像の濃淡値を I(x,y) とすると，局所領域 w におけるオプティカルフロー (u,v) は

$$u = \frac{\sum_w \frac{\partial I}{\partial x}\cdot[J(p) - I(p)]}{\sum_w (\frac{\partial I}{\partial x})^2} \quad (3)$$

$$v = \frac{\sum_w \frac{\partial I}{\partial y}\cdot[J(p) - I(p)]}{\sum_w (\frac{\partial I}{\partial y})^2} \quad (4)$$

さらに

$$I(p) = I(x, y, t) \quad (5)$$

$$J(p) = I(x, y, t, \delta t) \quad (6)$$

図 6, 7, 8 はこの方法で検出されたフローを描写したものである．


図 6　入力画像 1

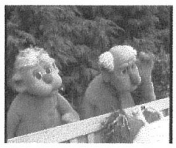
図 7　入力画像 2


図 8　結果画像

5.2.3 動作検出

Lucas-Kanade で検出されたオプティカルフローを利用してフレーム i と i+1 の動き情報 M_i を式 7 で定義する．

$$M_i = \frac{1}{N}\sum_{k=1}^{N}(x_{k,i} - x_{k,i+1})^2 + (y_{k,i} - y_{k,i+1})^2 \quad (7)$$

ただし N は見つかった全てのオプティカルフローの個数を，x，y はそれぞれの座標を表している．例えば $x_{k,i}$ は，i 番目のフレームの k 個めのフローの x 座標を表すものである．これは図 8 の線分の長さの平均を表す量である．この M_i が閾値以上であった場合それは連続するひとつの動作であるとみなす．

またショットとしての動き情報は Shot_M の式 8 で定義する．

$$Shot\_M = \frac{1}{F-1}\sum_{i=1}^{F-1} M_i \quad (8)$$

ただし F はショットに含まれる全てのフレーム数である．

5.3 顔認識

顔認識には openCV に含まれる顔検出プログラムルーチンを使用する[7]．k-means でクラスタリングした後，各クラスタの代表は顔情報による重み W とショットの動き情報を用いた式 9 で表される RE が最大のものが選ばれる．

$$RE = ALL\_MI \times W \quad (9)$$

この式によって，各クラスタからは顔が映っているものや動きが大きいものが優先的に選ばれるようにする．

6. 実　　験

6.1 ビデオデータ

実験では，TRECVID 2007 で用いられた development data 9 本を実験データとして使用した．ビデオの長さは最小 11 分，最大 36 分，平均で 21 分であった．

6.2 実験手順

評価するシステムとして以下の 3 つのシステムを考える

- 色情報のみを特徴量としたシステム
- 色情報と動き情報を特徴量としたシステム
- 色情報，動き情報，顔情報を特徴量としたシステム

評価基準として，いずれも TRECVID における公式な評価指標である，ground truth がどのくらいの割合で含まれているかを表す IN 値，オリジナルビデオの何%の要約になっているかを表す DU 値，システム実行にかかった時間 SYS 値を用いる．

最後にこの 3 つのシステムと TRECVID 2007 の参加者の結果を比較する．

6.3 実験結果

6.3.1 3 つのシステムの評価

表 2 から表 4 はそれぞれのシステムの結果を表している．ただし IN 値に関してはテキスト形式の ground truth との比較を主観的に比較するので評価者によって値が若干異なる場合がある．

表 2　色情報

	時間 [s]	IN	DU[%]	SYS[s]
rush01	2189	0.49	3.9	1488
rush02	2037	0.53	3.8	1386
rush03	721	0.61	3.7	613
rush04	738	0.38	10.4	1347
rush05	1951	0.63	3.8	1327
rush06	693	0.46	10.8	1348
rush07	743	0.62	3.7	525
rush08	767	0.42	9.5	1219
rush09	1702	0.66	3.8	592
平均	1282	0.50	4.3	1093

表 3　色情報+動き情報

	時間 [s]	IN	DU[%]	SYS[s]
rush01	2189	0.53	3.5	1999
rush02	2037	0.62	3.6	1608
rush03	721	0.46	3.7	864
rush04	738	0.50	7.7	1549
rush05	1951	0.60	3.5	1590
rush06	693	0.84	10.8	1719
rush07	743	0.75	3.7	657
rush08	767	0.36	5.4	1316
rush09	1702	0.66	3.6	742
平均	1282	0.55	4.4	1338

3 つのシステムとも rush08 の IN 値が比較的低くなっていることが分かる．これはこの動画が全体的に暗い画面であったことが原因であると考えられる．システムのなかで最も重要なクラスタリングは色特徴のみで行われているので，全体的に色特徴が変わらない動画においては精度が下ってしまう．

またどのシステムも ground truth の内容が”Shot

表 4　色情報+動き情報+顔情報

	時間 [s]	IN	DU[%]	SYS[s]
rush01	2189	0.62	3.6	2051
rush02	2037	0.60	3.3	1271
rush03	721	0.52	3.7	856
rush04	738	0.56	7.8	1540
rush05	1951	0.63	3.5	1590
rush06	693	0.76	10.8	1735
rush07	743	0.75	3.7	663
rush08	767	0.42	5.2	1298
rush09	1702	0.75	3.8	722
平均	1282	0.60	4.3	1433

of tree" のような内容については比較的良かったが，"Woman exit left" のような内容の場合，左を向いて退場する前に次の場面に切り替わってしまうことが起っていた．これは動き特徴を用いたシステムではある程度は改善されていたが，まだそのようなシーンは多く残っていた．

次に DU 値であるが，3 つのシステムともほとんどのビデオは規定の 4%以内に収まっているが rush04, rush06, rush08 において 4%を大幅に越えてしまっている．

システムごとの比較をしていく．まず特徴量が色情報のみのシステムと，色と動き情報を用いたものを比較する．IN 値において 0.50 と 0.55 と動きを用いた場合 5%の改善がみられた．このことから，動きを強調したものは精度が高くなることが分かる．ただしシステム時間においては，平均で 245 秒の差があり，計算コストがふえてしまっていることが分かる．

次に顔，色，動き情報を用いたシステムと，色，動き情報を用いたシステムの結果を比較する．これも IN 値は全体として 0.55 と 0.60 と 5%の精度が上がった．しかし一方で rush02 のように人間以外が中心となる動画の場合は顔特徴を用いない場合が精度が良いときがあった．システム時間に関しては，平均 95 秒の差があった．

6.3.2　TRECVID 2007 参加者との比較

図 9 は TRECVID の参加者との IN 値の比較である．ただし TRECVID の参加者の値は中間値であるが，本研究では用いたテストデータが少なかったので，式 10 で定義される値を用いた．

$$\frac{\text{全ての要約に含まれていた ground truth の数}}{\text{全ての ground truth の数}} \tag{10}$$

図 10 は DU 値の比較であるがグラフの横軸の値は式 11 で定義される．

$$\text{規定の要約時間 [秒]} - \text{実際の要約時間 [秒]} \tag{11}$$

よってグラフの負の値は製作した要約時間が規定の時間を越えていることを表している．

この 2 つの表において本研究の結果は上の 3 つである．ただし uec_all は特徴量として顔，動き，色を用いたもの，uec_c_m は色と動きを用いたもの，uec_color は色のみを用いたものである．

IN 値については比較的良い値が出せていることが分かる．しかし DU 値においては時間を越えてしまった 3 つの動画があったので，参加者と比べると，規定時間の超過が大きくなってしまっている．

7. 考　　察

図 11，12 はそれぞれ 4%以下の要約となったときの DU 値と IN 値の相関関係と 4% を超過したときの DU 値と IN 値の相関関係を示している．

図 11 から要約時間が短くなった場合は IN 値と DU 値は負の相関関係があることが分かる．更に図 12 からは逆に要約時間が規定を大きく越えた場合も負の相関関係より IN 値は低くなる傾向にあることが分かる．以上のことから要約時間を 4% になるように設定することは重要なことであると分かる．

表 5-7 はどのシステムにおいても 4%の時間を大幅に越えた動画に関してのそれぞれのシステムの IN 値の平均，その動画に関する IN 値，その動画に関する DU 値を表している．

表 5　Rush04

	IN 平均	IN 値	DU 値 [%]
C	0.50	0.38	10.4
C+M	0.55	0.50	7.7
C+M+F	0.60	0.56	7.8

表 6　Rush06

	IN 平均	IN 値	DU 値 [%]
C	0.50	0.46	10.8
C+M	0.55	0.84	10.8
C+M+F	0.60	0.76	10.8

表 7　Rush08

	IN 平均	IN 値	DU 値 [%]
C	0.50	0.42	9.5
C+M	0.55	0.36	5.4
C+M+F	0.60	0.42	5.2

ただし C，M，F はそれぞれ色，動き，顔特徴をあらわしている．実際に 4%を大幅に越えた動画については IN 値が同じシステムの IN 値の平均に比べて低くなる傾向があることが分かる．

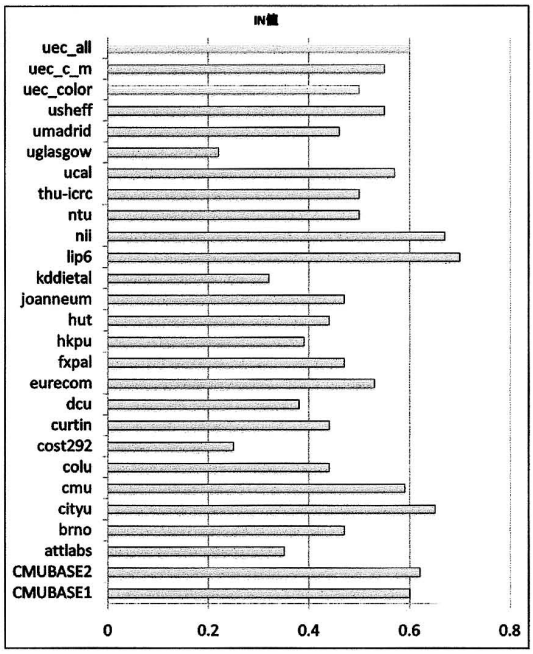

図 9 IN 値

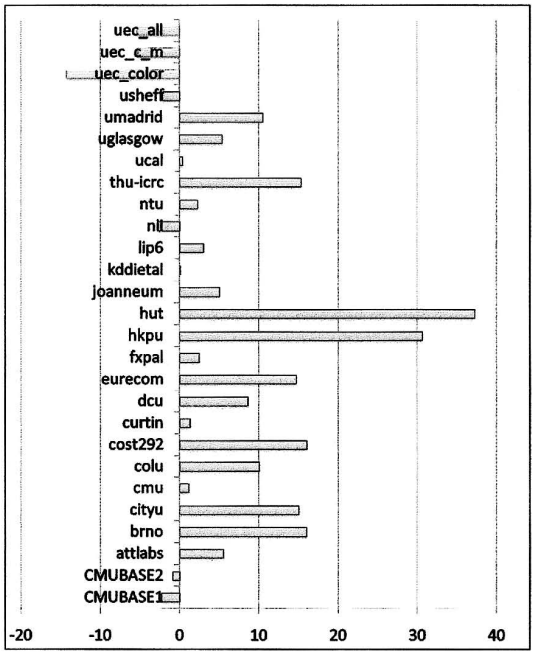

図 10 DU 値

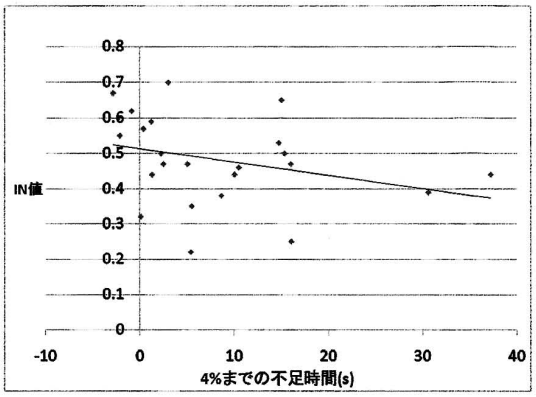

図 11 IN 値と DU 値の相関関係 1

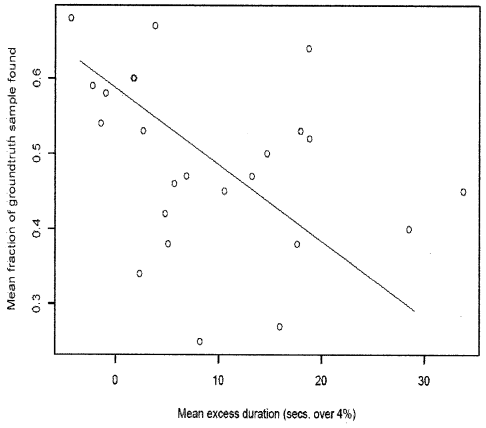

図 12 IN 値と DU 値の相関関係 2

またジャンクショットの検出もヒストグラムインターセクションでおこなっているが図 13 に示すカラーバーは右は検出することができたが，左の画像は検出することができなかった．これは動画によってカラーバーの明るさが微妙に違ってくるからである．これではこの方法では検出することができない．よってジャンクショット検出のさいは EMD(Earth Mover's Distance)[8] などの他の手法を考えなくてはいけない．

本システムではクラッパーボードの検出についてなにも行っていない．このことで要約中にジャンクショットが含まれていた．更にクラッパーボードが残ってしまっているために tiny shot 分解の際の特徴量がしっかりとれなくなり，システムの精度を下げてしまっている．

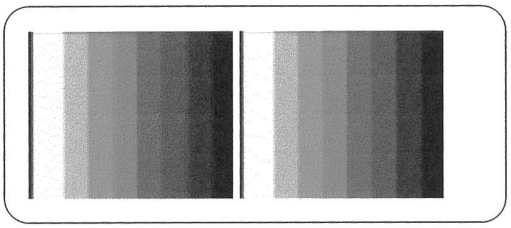

図 13 カラーバー

しかしこれを視覚特徴のみで検出することは極めて難しい．その理由として，クラッパーボードの形の多様性，画面のどの場所に現れるかは決まっていないことが挙げられる．図 14 はクラッパーボードの現れ方の違いを表している．

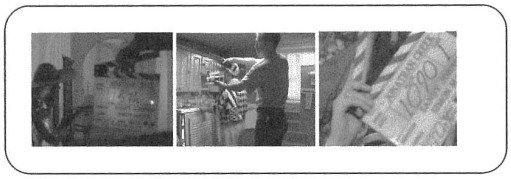

図 14 クラッパーボード

4) では音響エネルギーが急激に増加することを利用した検出方法を提案している．本研究ではこの音特徴と視覚特徴を組合わせた手法を取り入れていく．ここで音特徴のみで行わないのは，それだけでは人が叫ぶシーンや，事故のシーン，車などのドアが閉まるシーンも検出してしまうからである．

8. おわりに

8.1 まとめ

本研究では，映像自動要約タスク (rushes summarization) について取り組んだ．その手法として，色特徴をもとにクラスタリングされた細かいショットを動き，顔特徴をもとに重みづけを行い，クラスタ代表を決定，つなぎあわせる手法を提案した．

まず用いる特徴量を変えた 3 つのシステムについて評価を行い，それらを比較した．その結果として色，顔，動きの 3 つの特徴量をもちいたシステムが最も良い動作を得ることが分かった．ただし特定の動画にたいして規定の時間である 4%を大幅に越えてしまった．

次に，これらのシステムと TRECVID 2007 の参加者との比較を行った．結果として IN 値については参加者と比べ良い結果であった．ただし DU 値については，いくつかの動画に関して，4%を大幅に越えたものがあったので，参加者と比べて，要約時間が長くなってしまった．

8.2 今後の課題

まず要約時間が 4%を越えた原因を明らかにし，改善していく．そしてそのことがシステムの精度にどのような影響を与えるかを検証していく．

次にカラーバーなどのジャンクショット検出について EMD などの新たな手法を取り入れる．そしてその精度を評価し，本システムとの比較を行う．更にクラッパーボード検出のために音特徴を取り入れていく．

また現在使われている特徴量についても，改善し，そのシステムの評価をしていく．例えば色特徴は現システムでは RGB 色空間を用いているが，HSV 色空間を取り入れることによるシステムの影響を調べていく．顔特徴も現システムでは正面顔のみであるので，これを横顔などの検出も行うシステムを組み込み，その評価を行っていく．

参考文献

1) P.Over, A.Smeaton, and P.Kelly. The trecvid 2007 bbc rushes summarization evaluation pilot. In *Proc. of the international workshop on TRECVID video summarization*, pp. 1–15, 2007.
2) D.Le and S.Satoh. National institute of informatics, japan at trecvid 2007: Bbc rushes summarization. In *Proc. of the international workshop on TRECVID video summarization*, pp. 70–73, 2007.
3) A.Hauptmann, M.Christel, W.Lin, B.Maher, J.Yang, R.Baron, and G.Xiang. Clever clustering vs. simple speed-up for summarizing rushes. In *Proc. of the international workshop on TRECVID video summarization*, pp. 20–24, 2007.
4) Y. Liu, Y. Liu, and Y. Zhang. The hong kong polytechnic university at trecvid 2007 bbc rushes summarization. In *Proc. of the international workshop on TRECVID video summarization*, pp. 50–54, 2007.
5) B.Lucas and T.Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. of International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.
6) B.Horn and B.Schunck. Determining optical flow. *Artificial Intelligence*, pp. 185–203, 1981.
7) OpenCV. http://opencv.jp/.
8) Y.Rubner, C.Tomasi, and L.J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, Vol.40, No.2, pp. 99–121, 2000.