クラスタリングによるTRECVIDラッシュ映像の要約*

野口 顕嗣　柳井 啓司
電気通信大学 情報工学科†

1 はじめに

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．具体的方法としては，[1] を参考にして，ショット分割した映像を色，動き，顔特徴に基づいてクラスタリングすることによって，映像要約を実現する．そして，実験によって，その有効性を示す．

2 TRECVID について

TRECVID とは映像コーパスを用いた情報検索のための競争型ワークショップで米国の NIST(National Institute of Standards and Technology) の主催で行われている．その主な目標はビデオの content-based 検索の向上である．

TRECVID 2007 において以下の 4 つのタスクが設定された．

- Shot boundary detection(ショット境界検出)
- High-level feature extraction(高次元特徴抽出)
- Search(検索)
- Rushes summarization(ラッシュ映像要約)

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

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．

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 [4]．

図 1，2 はそれぞれラッシュ映像の 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: ラッシュ映像の 10 秒毎のフレーム

3 アルゴリズム

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 3 はシステムの概要を表している．

最初に与えられたビデオを色特徴をもとに，前後のフレームを比較し色ヒストグラムの差分が閾値以上なら

図 2: 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

ショット分割する．また各ショットの色特徴からクラスタリングを行いそれぞれのクラスから最も長いショットを代表として選んでいく．

その際，ブラックフレームや，カラーバーなどのジャンクショットの検出をクラス単位で色特徴を用いて行う．

このようにして得られた各クラスタの代表から今度は色特徴，顔特徴，動き情報を抽出しながら，各クラスタの代表をそれぞれ一秒単位に分割する，ただしこの際に Lucas-Kanade 法 [2] でオプティカルフローを計算して，ある一定以上の動きがあった場合にそれ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，分割を行わないようにする．

これにより [1] においてはカメラモーションのみの強調であったが，本研究では動作も強調できるようにする．

その一秒毎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．

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ほしいので，動き情報と顔情報を用いてクラスタの代表を決定する．各クラスタの代表を時間順につなぎあわせて，要約映像とする．

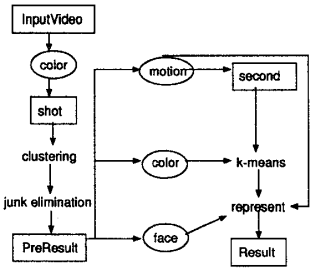

図 3: アルゴリズム概要

4 特徴量

本章ではショットを表す特徴量の算出方法について説明する．

4.1 位置情報付カラーヒストグラム

色特徴としては 3 × 3 に分割した RGB カラーヒストグラムを使用する．各ショットの色特徴 CF は式 1 で定義する．

*TRECVID Rush Video Summarization by Clustering
†Akitsugu Noguchi and Keiji Yanai, Department of Computer Science, The University of Electro-Communications ({noguchi-a,yanai}@mm.cs.uec.ac.jp)

$$CF = (\sum_{i=1}^{F} I_i)/F \quad (1)$$

ここでFはそのショットのフレームの枚数．Iはそれぞれのフレームにおけるカラーヒストグラムを表している．

4.2 動き情報

Lucas-Kanade法 [2] によって連続するフレーム間のオプティカルフローを計算する．2つのフレーム間における動き情報MIを式2で定義する．

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

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．

そしてショットとしての動き情報，ALL_MIは式3で定義される．

$$ALL\_MI = (\sum_{k=1}^{F} MI_k)/F \quad (3)$$

ただしFはビデオに含まれる全てのフレームの数である．

4.3 顔特徴

顔の認識はOpenCV[3]の顔画像検出プログラムルーチンを利用する．また顔が検出されたものには，重みW(実験ではW=1.5)をつける．ただし顔が検出されなかった場合W=1である．k-meansでクラスタリングしたあとに各クラスタの代表を次のREの値が最大のもので定義する．

$$RE = ALL\_MI \times W \quad (4)$$

5 実験

5.1 実験データ

実験データとして，TRECVID2007で，提供された発展データを使用する．この実験で用いたビデオは，全部で9本であり，ビデオの長さは最大36分，最小11分，平均値は，約21分であった．

5.2 評価基準

ここではこのシステムについてTRECVIDの評価方法の中から4つの評価基準を用いる．一つめはそのビデオに対するground truthがどれだけの割合で含まれているかを表すIN値，2つめはオリジナルビデオに対して何パーセントの要約になっているかを示すDU値，3つめがシステムの実行にかかった時間を表すSYS値である．

5.3 実験結果

結果は表2で示すようになった．

実験でのIN値の平均は0.60となった．TRECVID 2007の参加者でIN最大のチームのメディアンが0.70で，中間のチームが約0.50であった．

ただしこの評価は評価者が要約映像とテキストで記述されたground truthが一致するかを主観的に評価したものであり，評価者によって値が若干異なる可能性がある．

個々の結果を見ていくと，まずIN値であるがほとんどが0.6前後であることが分かる．しかしrush08のIN値が低くなってしまっている．この原因としてて考えられることは，このビデオが全体を通して黒中心の色をしていたからである．クラスタリングを行う際は色で行っているので，全体的に似た色の場合精度が下ってしまっている．

[1]のIN値は0.59となっている．これは[1]に改良を加えた本研究とほとんど変わらない結果となっている．原因としては[1]で実装されていクラッパーボードなどのいくつかのジャンクショット検出が本研究では未実装であること，動き情報の検出が不完全であることが考えられる．

最後にSYS値は，TRECVIDの参加者の多くが1000以下だったことに比べて若干長くなってしまっている．

表2: 評価のまとめ

	時間[s]	IN	DU[%]	SYS[s]
rush01	2189	.0.62	3.6	2051
rush02	2037	0.60	3.3	939
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	1066

6 考察

実験結果よりこのシステムの欠点がいくつか分かった．第一に実行にかかる時間が比較的長いことが挙げられる．その原因として挙げられることが，色情報を抽出する作業がショット検出とk-meansの特徴とで重複していることである．

二つ目としては，要約として，見難くなっている．一秒毎に，場面が切り替わってしまうので，みている側も理解することが大変になっている．

またIN値において，ground truthの内容が"Shot of tree"のようなものは精度が比較的高かったが，"Woman exit left"のような内容のとき，左に行く途中で次の場面に切り替わってしまう部分が多かった．これは動き情報が不完全であることを示している．

更にクラス分けは色を中心に行っているので，rush08のように暗い場面が大半をしめている動画に関して，精度が下ってしまっている．

7 おわりに

本研究ではラッシュ映像の要約の方法として提案された[1]の手法に対して動き，顔情報といった新たな特徴を加えた手法を提案しその結果について述べた．

今後の課題として，クラッパーボードのようなジャンクショットの検出を実装すること，ほかに動き情報の改良などの，特徴量の改良，音情報などの現在未使用の特徴量の追加が挙げられる．

参考文献

[1] A.G. Hauptmann, M.G. Christel, W.H. Lin, B. Maher, J. Yang, R.V. Baron, and G. Xiang. Clever Clustering vs. Simple Speed-Up for Summarizing BBC Rushes. *Proc. the TRECVID Workshop on Video Summarization*, pages 20–24, 2007.

[2] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *Proc. International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[3] OpenCV. http://opencv.jp/.

[4] P.Over, A.F.Smeaton, and P.Kelly. Trecvid 2007 bbc rushes summarization evaluation pilot. *Proc. the TRECVID Workshop on Video Summarization*, pages 1–15, 2007.