I-005

Bag-of-Visual Words 表現を用いた放送映像中の類似シーン検出
Similar Scene Detection in Broadcasted Videos using Bag-of-Visual Words

井尻 将太 Shota Ijiri　貞元太志 Taishi Sadamoto　黒木 修隆 Nobutaka Kuroki　廣瀬 哲也 Tetsuya Hirose　沼 昌宏 Masahiro Numa

1. はじめに

近年，放送のディジタル化や大容量 HDD レコーダの普及によって視聴者が蓄積する映像情報が増加している．それらの視聴を支援するため，映像をシーンごとに分割・整理することによって構造化する研究が行われている．

映像を構造化する手法として，スポーツ映像に特有の規則性を用いる手法 [1]，文字テロップを利用する手法 [2] などが存在する．しかし，これらの手法は適用可能なシーンが限定されている．

本研究では，映像内の画像特徴とその出現時刻の情報を用いることで汎用性の高いシーン分割を行う．

2. 映像内のシーン範囲推定

提案手法では Bag-of-Visual Words と呼ばれる画像認識手法を用いて映像内の類似フレーム検出を行い，それらの出現時刻を用いてシーン範囲の推定を行う．

2.1 Bag-of-Visual Words(BoVW)

BoVW は局所特徴と呼ばれる画像内の部分的な特徴ベクトルの出現頻度を用いた物体認識技術であり，その特徴表現の生成は，

i) 学習用画像から局所特徴量を抽出

ii) 特徴量のクラスタリングによって Visual Words を生成

iii) 検出対象画像における Visual Words の出現頻度表(ヒストグラム)を作成

の三段階からなる．以下に詳しい手順を述べる．

2.1.1 局所特徴量の抽出

BoVW の第一段階として学習用画像から局所特徴量を抽出する．本手法ではこの局所特徴量に SURF と呼ばれる特徴を使用する．SURF の処理は特徴点抽出と特徴記述の二段階に分かれており，この特徴量は局所的にコントラスト変化の大きい点を特徴点として抽出する．

2.1.2 Visual Words の生成

Visual Words 生成のイメージを図 1 に示す．画像上の円は抽出した特徴点を示している．これらの特徴量を k-means 法によりクラスタリングする．クラスタリングによって生成された各クラスタの中心を Visual Words と呼ぶ．

2.1.3 BoVW 表現のヒストグラム作成

検出対象画像から SURF を抽出し，それらを最も近い Visual Words に投票する．これにより画像内に存在する Visual Words のヒストグラムを表すことができる．映像内の各フレームにおいてこのヒストグラムを作成する．局所特徴量を BoVW 表現に変換することは，局所特徴量を量子化しているといえる．

2.2 類似フレームの検出

映像内の 1 枚のフレームを検索の元となる画像（クエリ）とし，全てのフレームに対して類似フレーム検出を行う．クエリフレームと他の各フレームの BoVW 表現のヒストグラムの類似度 R を算出し，図 2 のように閾値を上回るフレームを類似フレームとする．

2.3 類似シーン範囲の推定と構造化

検出された類似フレームが時間的に連続して存在する部分を図 3 のようにシーン範囲として出力する．出力されたシーンの中で，クエリフレームが含まれるシーンをクエリシーンとすると，出力された他のシーンはクエリシーンの類似シーンとなる．このように類似シーンの情報を各シーンに付加することで映像を構造化する．

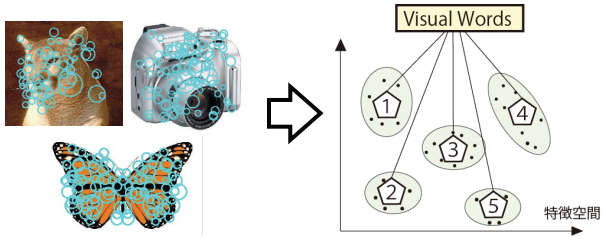


図 1　Visual Words の生成

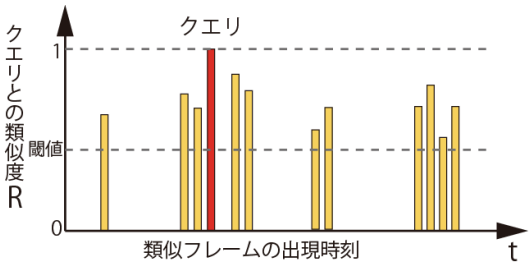


図 2　類似フレームの検出結果

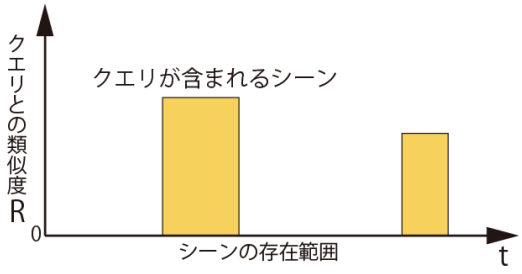


図 3　シーン範囲推定結果

表 1　ニュース番組のシーン範囲推定結果(%)

シーン	適合率	再現率	F 値
Map	100.0	75.2	79.5
Baseball	71.6	68.1	54.7
Interview	69.9	100.0	79.6
Golf	100.0	51.2	63.5
平均	85.4	73.6	69.3

表 2　バラエティ番組のシーン範囲推定結果(%)

シーン	適合率	再現率	F 値
Studio1	97.9	95.1	96.1
Location1	95.8	97.3	96.2
Location2	95.2	49.2	55.1
Location3	88.1	24.9	56.1
平均	94.3	66.6	70.3

表 3　音楽番組のシーン範囲推定結果(%)

シーン	適合率	再現率	F 値
Talk1	36.2	90.5	48.8
Talk2	77.3	92.3	80.7
Sing1	100.0	72.6	78.4
Sing2	100.0	83.3	88.6
平均	78.4	84.7	74.9

3. 評価実験と考察

3.1 評価実験

実際に放送されたニュース，バラエティ，音楽の 3 ジャンルのテレビ番組映像に対して，クエリシーンの範囲を推定する実験を行った．内容的に連続するシーンを予め人手によって定め，各ジャンルから 4 シーンずつ実験対象として選出する．各シーンからランダムに抽出したクエリフレームを元に類似フレームの検出，シーン範囲の推定を行う．

出力したクエリシーン内の各フレームと，正解シーン内の各フレームの一致率によって推定精度を評価する．評価指標としては，出力フレームの正確性を表す適合率，正解フレームの網羅性を表す再現率，それらの調和平均である F 値を用いる．

3.2 実験結果と考察

シーン範囲推定の評価結果を表 1，表 2，表 3 に示す．各シーンに対してクエリは 10 枚抽出し，それぞれの推定精度の平均値を求めた．平均すると F 値は 70%を少し上回る程度である．クエリ毎の F 値を見ても 80 %以上の高い値が多く見られる．これはほとんどのクエリで正確なシーン範囲の推定が出来ていることを示す．音楽番組におけるシーン「Sing1」で用いたクエリフレームと，その類似フレーム検出結果の一部を図 4 に示す．図 4 では BoVW を用いたことによって，部分的な類似点が多いフレームを類似フレームとして検出することができている．バラエティ番組の「Studio1」では，シーン内のフレームの構図がほぼ等しいものが多かったため，非常に高い精度で検出できた．

しかし，結果の中には再現率のみが極端に低いシーンが存在する．例えばニュース番組の「Golf」ではシーン内で構図が大きく変化したため，それらを 1 つのシーンとして検出できなかった．またバラエティ番組の「Location3」では，図 5 のように背景から抽出される局所特徴が少なく，文字・テロップの有無が BoVW 表現のヒストグラムに大きく影響し，類似フレーム検出がうまく行えなかった．これらの原因から再現率が低下した．

4. まとめ

本論文では，放送映像の構造化を目的として，BoVW を利用した類似フレーム検出を行い，その検出結果を用いてシーン範囲を推定する手法を提案した．

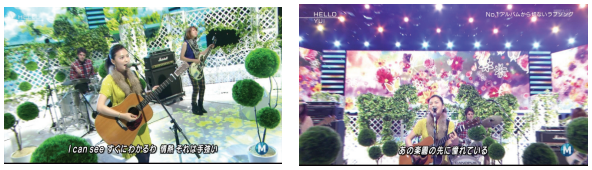
今後の課題としては，類似フレームの取りこぼしによるシーンの過分割を防ぐことが挙げられる．

参考文献

[1]椋木雅之,寺尾元宏,池田克夫, “カット構成の規則性を利用したスポーツ映像のプレイ単位への分割”,電子情報通信学会論文誌 D-Ⅱ,Vol.J85-D-Ⅱ,No.6,pp.1016-1024，(2002).

[2]三浦宏一,高野求,浜田玲子,井手一郎,坂井修一,田中英彦, “料理映像の構造解析による調理手順との対応付け”,電子情報通信学会論文誌 D-Ⅱ,Vol.J86-D-Ⅱ,No.11,pp.1647-1656,(2003).

(a) クエリフレーム



(b) 類似フレーム

図 4　類似フレームの検出結果例

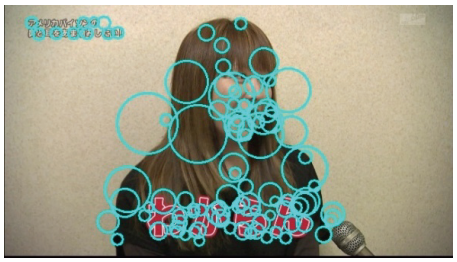


図 5　Location3 における特徴点