H-012

H.265 圧縮動画における特定動作と人物の特徴量空間上での分類

高場 雄太　森田啓義

電気通信大学
t2031095@edu.cc.uec.ac.jp, morita@uec.ac.jp

1　はじめに

現在、映画、ドラマ、スポーツ中継、アニメなどといった映像作品は、携帯端末で視聴されることが一般的となっている。この携帯端末向けのデータ圧縮技術として用いられているのが H.265[1] である。現在地上デジタル放送などで使用されている標準的な動画圧縮技術 MPEG-2[2] と比較して多くの符号化情報が含まれており、これらの情報を用いることによって、MPEG-2 で符号化した場合に比べ符号後のデータ量を $\frac{1}{4}$ に削減できるという高能率符号化を実現している。これは、符号化に際して H.265 は他の方法に比べ、より高度な動画解析を実施している結果であると言える。

動画解析は、一度符号化された動画を復号し、そこから動画の解析を行うというのが一般的な解析方法である。しかし、前述したように、H.265 で圧縮し符号化された動画データには、高度な動画解析がすでに実施されているので、その情報を用いることで、完全な動画復元を行わずとも動画の解析ができるのと考えられる。また、符号化パラメータを用いて動画の解析を行った研究として、動きベクトルの情報を用いて動体検出を行ったもの [3][4] が報告されている。上記の研究を参考にし、完全な動画復元を行わず、符号化パラメータから動画の特徴を抽出すれば、動画内の人の動作や人物の分類が可能になると考えた。

本稿では H.265 で圧縮された動画の符号化情報から、符号化単位である CU の大きさや、動きベクトルの大きさというような動画の特徴を表す符号化情報を抽出し、その抽出した符号化情報から新たな特徴量を生成した。その生成した特徴量から、各動作、各人物の特徴を学習した線形モデルを構築した。そのモデルを用いて動画内における人物の歩く、手を振るなどといった動作がその符号化情報のみで分類できることを示す。

2　提案手法の概要

2.1　提案手法の流れ

H.265 で圧縮された動画の符号化情報から、動作と人物の分類を行う流れを以下に示す。

1. 符号化パラメータから特徴量の作成
2. 作成した特徴量を用いて SVM[6] を使い学習を行う
3. 学習を行ったモデルを用いて分類を行う

2.2　H.265 の符号化パラメータ

H.265 で符号化された動画の符号化情報のうち、本研究で特徴量として使用したものは以下の情報である。

- 「I」「P」「B」いずれかのスライスタイプ: S_t
- 予測に参照画像を用いるかどうかを示すフラグ情報: F_i
- CU サイズ (8 画素〜64 画素): C_s
- 予測タイプ (イントラ予測: 0, インター予測: 1): P_t
- 動きベクトルの x 方向、y 方向の値: x_1, y_1, x_2, y_2

上記の 8 つの CU の情報から、以下のような 8 次元の特徴量を作成した。

$$\boldsymbol{V} = (S_t, F_i, C_s, P_t, x_1, y_1, x_2, y_2) \tag{1}$$

上記にある CU とは、画面をいくつかのブロックに分けたもののことである。この CU をいくつか組み合わせ、固定サイズのブロックとしたものを CTU といい、1 つ以上の CTU の集合からなるものをスライスという。

また、動きベクトルとは、各 CU において前後の動きとの差分を示したものである。動きベクトルに関しては、参照画像を最大二つまでとることができ、参照画像が存在しない場合は、0 とした。

3　提案手法

3.1　特徴量の作成

特徴量の作成方法と分類方法について詳しく記述する。前節の式 (1) で与えた $\boldsymbol{V}$ は CU 単位での特徴量である。CU の数はピクチャ毎に異なっている。ピクチャ単位に特徴量をまとめたものを処理したいので、ピクチャ単位の特徴量の次元を揃える必要がある。特徴量の次元数を揃えるために今回は以下の 2 つの方法を取った。

方法 1　重みをつけた特徴量の作成

方法 2　0 埋めを行った特徴量の作成

3.1.1　重みをつけた特徴量の作成

サイズが小さい CU には、動きの特徴が多く含まれていると考えらるので、小さい CU の情報をより強く抽出するために、$\boldsymbol{V}$ について重み付けを行った。これは、サイズの大きい CU の特徴量を最小の CU サイズである 8×8 画素サイズに分割して特徴量とするための重みづけの方法である。

以下のように、重みづけを行った。

$$\boldsymbol{V} = \boldsymbol{V} \times (\frac{8}{C_s})^2 \tag{2}$$

3.1.2　0 埋めを行った特徴量の作成

この方法は、CU の左上の座標にその CU 情報を付与し、その他の部分は特徴量を 0 とする方法である。

具体的には、以下のような計算式を用いて、特徴量の作成を行った。ここで、CU の左上の x 座標を CU_x、左上の y 座標を CU_y とする。

$$\boldsymbol{V} = \begin{cases} \boldsymbol{V} & \text{if} \quad \text{x 座標} : \mathrm{CU_x}, \text{y 座標} : \mathrm{CU_y} \\ 0 & \text{else} \end{cases} \tag{3}$$

3.2 SVM を用いた学習

3.1 で作成した、ピクチャ単位で大きさを揃えた特徴量に対して、30 フレーム内の移動平均の列を SVM への入力とした。

ある人物がある動作を行う動画毎に特徴量を作成し、人物と動作についてそれぞれラベル付けしたものを訓練データとして使用した。この作成した訓練データを元に、動作で分類を行うモデル、人物で分類を行うモデルをそれぞれ SVM で学習した。SVM は scikit-learn[5] を使用した。

3.3 学習済みモデルを用いて分類

新たな動画から訓練データと同様に 2 つの方法で特徴量を作成した。3.2 で学習したモデルにその作成した特徴量をテストデータとして入力し、各動画ごとに人物や動作の検出を行った。その検出結果から、分類が可能であるかの検証を行った。

4 実験結果

実験に使用した動画は、中央に 1 人の人が写っており、その人物が「ボクシング」「歩く」「手を上下に振る」「手を叩く」のいづれかの動作を行っているというものである。動画の長さは約 5 秒程度で、訓練データとして使用した動画の総数は 24 個である。テストデータとして使用した動画は、訓練データとして使用したものと同様の形式の動画になっており、総数は 4 個である。

4.1 分類可能性の検証

まず、訓練データに含まれる 3007 個の特徴量を 8:2 に分割し、2405 個の特徴量を用いてモデルの学習を行い、残りの 602 個の特徴量を用いて、F 値の出力を行った。検証実験は 5 回繰り返した。毎回、ランダムに訓練データの分割を行い、結果の平均を求めた。表 1 にその結果を示す。

また、テストデータを用いて実際に入力動画に対して、動作と人物の分類ができるのかの確認を行った。動画における人物と動作の導出方法を以下に示す。

3.2 より、1 本の動画から、フレーム数 - 30 個の特徴量が取り出せる。取り出した特徴量ごとに人物と動作の検出を行い、最も多く検出された人物と動作を選出した。

表 2 に 1 本の動画から得られた特徴量における各動作の検出割合と、正解となる人物と検出できたものの割合をそれぞれ示す。

5 考察

表 2 に示すとおり、手を振る動作以外で 70%前後の正答率が得られた。手を振る動作において、正答率が低くなっているのは、他の 3 つの動作と比較すると動作における特徴的な動きの時間が局所的であり、動き自体の特徴が小さくなっていることが原因と考えられる。

重みづけを行った特徴量と、0 埋めを行った特徴量では、全体的に 0 埋めを行った特徴量で実験を行った方が正答率が良いことがわかる。その理由は、重みづけを行っている特徴量と比べて、各特徴量の大きさを小さくしなかったことにより、動作や人物の特徴がよりよく出たのではないかと考えられる。

しかし、手を振る動作では重みつけを行ったものの方が正答率の高い結果となった。このことから、重みつけを行いサイズが小さい CU の情報を強く抽出することで、局所的な時間における動きの検出の精度が高くなるのではないかと考えられる。

6 まとめ

本研究では H.265 で圧縮された動画の符号化情報から、CU や、動きベクトルの大きさというような動画の特徴を表すような符号化情報を抽出し、そこから新たな特徴量を生成し、人物の歩く、手を振るなどといった動作がその符号化情報のみで分類できることの検証を行った。

結果として、動画の分類自体は可能であったが、各動画における正答率については課題の残る結果となった。手を振る動作のような局所的な特徴もうまく学習することができるような新たな特徴量の作成方法の検討を行っていき、実世界のデータセットへの拡張も視野に入れ、今後の研究を行っていきたいと考えている。

分類区分	重みづけ	F 値
動作	重みづけ	0.741
人物	重みづけ	0.738
動作	0 埋め	0.858
人物	0 埋め	0.830

表 1: 特徴量の作成方法と F 値の関係

動作内容	方法	動作検出率 (b, c, w, s)(%)	人物検出率 (%)
ボクシング (b)	1	**72.8**, 26.4, 0.8, 0	52.8
ボクシング (b)	2	**67.2**, 32.0, 0.8, 0	70.4
手を叩く (c)	1	8.8 , **76.0**, 4.0, 11.2	74.4
手を叩く (c)	2	16.0, **78.4**, 4.8, 0.8	83.2
手を振る (w)	1	17.6, 30.1, **39.0**, 13.3	58.0
手を振る (w)	2	23.5, 31.6, **40.4**, 4.5	54.4
歩く (s)	1	12.2, 10.7, 3.3, **73.8**	62.3
歩く (s)	2	15.5, 3.3, 2.5, **78.7**	72.1

表 2: 各動画における動作と人物の検出率

参考文献

[1] 大久保 榮監修, 鈴木 輝彦 編, 高村 誠之 編, 中條 健 編, ”インプレス標準教科書シリーズ H.265/HEVC 教科書”

[2] 藤原洋, ”最新 MPEG 教科書”, アスキー社出版局 (1994)

[3] Yung-Wei Chen, Kai Chen, Shih-Yi Yuan, Sy-Yen Kuo, ”Moving Object Counting Using a Tripwire in H.265/HEVC Bitstreams for Video Surveillance”, IEEE(2016)

[4] Serhan Gül, Jan Timo Meyer, Cornelius Hellge, Thomas Schierl, Wojciech Samek, ”Hybrid video object tracking in H.265/HEVC video streams”, IEEE(2017)

[5] scikit-learn, `https://scikit-learn.org/`

[6] Vapnik, V. ”The Nature of Statistical Learning Theory, Springer”, New York(1995)