I-033

映像符号化データを用いた Two-stream CNN による映像認識の検討
A Study of Action Recognition Based on Two-stream Convolutional Networks using Compressed Video Data

築地　渉太†　　八島　由幸†
Shouta Tsukiji　Yoshiyuki Yashima

1. はじめに

動画認識は，一般的に空間情報と時間情報を組み合わせて行われる[1][2]．代表的な手法の 1 つに Two-stream CNN がある[2]．CNN への入力として，空間情報をとらえるための RGB 値と，動画の時間情報をとらえるためのオプティカルフローが用いられるが，オプティカルフローは算出コストが高いという欠点が存在する．一方で動画像は，通常 H.265/HEVC などの高効率な圧縮符号化データとして扱われる．本研究では，圧縮ビットストリームに含まれる動きベクトル情報を時間情報として，DCT 係数情報を空間情報として利用する Two-stream CNN を検討する．使用する DCT 係数の周波数成分や，エンコードの際の量子化パラメータによる認識率の違いを考察し，従来の Two-stream CNN との認識精度の比較を行う．

2. 部分復号情報を用いた学習

2.1 処理の流れ

図 1 に認識の流れを示す．本検討では認識部分に対して Two-stream CNN を利用し認識を行う． Two-stream-CNN は，入力として時間方向に重ねられたオプティカルフローと RGB 画像を用いて各 CNN を学習させ，それぞれの出力を融合させる． RGB 画像に対してオプティカルフローが相補的な特徴をとらえているため認識率が向上する．しかし，問題点としてオプティカルフローの計算コストの高さがあげられる．本検討では HEVC 符号化データに含まれる動きベクトルや DCT 係数のみから認識を行うことで計算コストを下げながら認識率を維持する方法を検討する．

2.2 動きベクトルと補完

本検討では動画認識によく用いられるオプティカルフローに代わり動きベクトルを用いる．動きベクトルは H.265/HEVC において符号化の際に利用する情報であるため映像データから抽出が可能である．しかし，動きベクトルを使用するデメリットとして，動きベクトルの信頼性，解像度の低下，イントラ予測領域のデータの欠落といった問題がある．そのため本検討では，まず，イントラブロックの部分に対して，周囲の動きベクトルを利用したメディアンフィルタを適用したベクトル補完を行う．その後信頼性の低い動きベクトルを除去する目的で全ブロックに対してメディアンフィルタを施す．こうして完成した水平方向と垂直方向の動きベクトルデータに対し平均減算を行い，図 2 のように交互に重ねたものを学習のデータとして利用する．

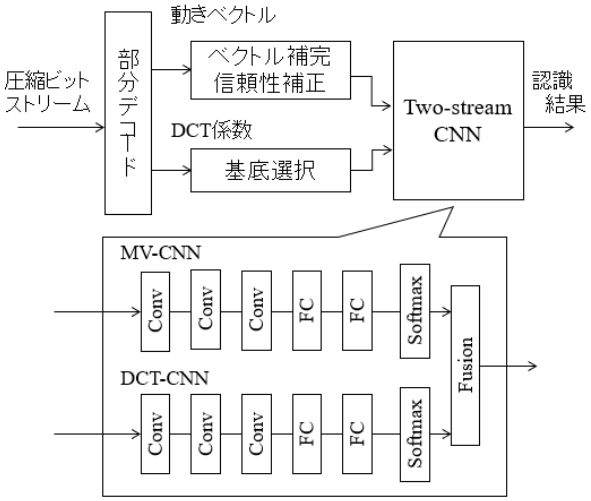

図 1　処理の流れ

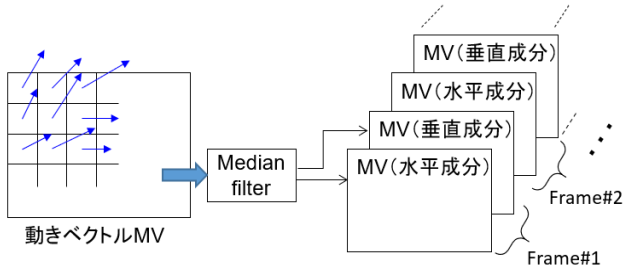

図 2　動きベクトルの CNN 入力

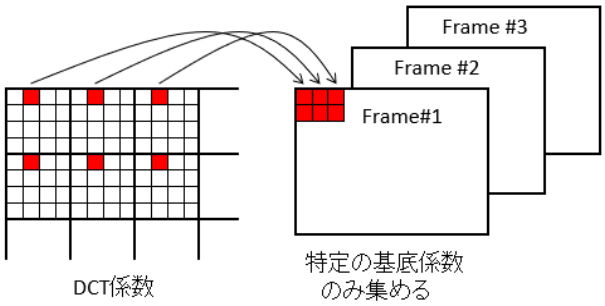

図 3　DCT 係数フレームの CNN 入力

† 千葉工業大学大学院情報科学研究科，Graduate School of Information and Computer Science, Chiba Institute of Technology

2.3 DCT 係数

DCT とは信号情報を周波数領域へ変換する手法であり，人間の視覚特性として低周波成分の変化に敏感であるという特性を用いて圧縮を行っている．HEVC によるエンコード時には，動き補償予測誤差情報に対してブロック（Transform unit, TU）ごとに 2 次元 DCT を適用することで DCT 係数情報が算出され，量子化されてビットストリームに埋め込まれる．デコード側ではビットストリームを部分デコードして得られる量子化 DCT 係数を 8bit 画像情報として利用できるように丸め込みを行う．その後，図 3 に示すように，定められた基底成分に対応する DCT 係数のみを集めてフレームを構成する．TU サイズを N×N とすると，DCT 係数フレームの大きさは，もとの画像サイズの 1/N×1/N となる．DCT 係数フレームを時間方向に並べて CNN の入力とする．

3. 実験結果

3.1 実験条件

実験には UCF101 データセットから 10 クラスを利用する[3]．解像度は 320 x 240 である．HEVC エンコードパラメータとして CTU サイズ 16，CU サイズ 8，TU サイズ 8 で固定し，動きベクトル探索領域±16，探索精度 1/4 とした．量子化パラメータ QP は 16，27，37 に設定した．ブロックサイズが 8 であるため CNN 入力の動きベクトル，DCT 係数フレームの解像度が水平垂直とも 1/8（40×30）になる．入力フレーム数は 20 フレームとした．動きベクトルでは 2.2 で記した補完処理を行った後，水平方向成分フレームと垂直方向成分フレームを交互に入れるため 40 フレームとなる．本実験では，動きベクトル（MV），DCT 係数直流成分（DCT16, DCT27, DCT37，数値は QP 値を示す）を CNN 入力とするほか，比較対象の従来技術として，原画像の輝度成分（Y）とオプティカルフロー（OF）を入力対象とした場合の実験も行った．Two-stream CNN の統合手法としては平均化を用いた．訓練時の損失関数は Closs Entropy Loss，最適化関数として SGD を用い，学習率 0.01，モーメンタムを 0.9 として学習を行った．評価指標はテストデータに対する認識率とした．

3.2 実験結果と考察

実験結果を図 4 に示す．図 1 の(a)～(f)は各データを単独で用いた場合の認識率を示す．動きベクトル単独ではオプティカルフローよりも空間解像度が粗いことから，認識率は低下することがわかる．また DCT 係数の量子化は粗くなるに従って認識率が低下するが，輝度信号値そのもので認識する場合に比べると，QP=16 では認識率はかえって向上するという結果になった．

図 4 の(g)～(i)は Two-stream CNN を適用した場合の認識率を示す．単独データを入力としたいずれの場合より認識率が向上していることが確認できる．また DCT 係数（DCT16）と動きベクトル（MV）の Two-stream CNN は，従来の輝度信号（Y）とオプティカルフロー（OF）の Two-stream CNN に比べてそん色ない認識率を達成できていることがわかる．図 5 の混同行列を見ると，クラス 1 やクラス 9 のように，従来よりも大きく正解率が向上するような動作映像もあることが確認できる．このことから圧縮ストリームを完全デコードしてから従来の Two-stream CNN を用いなくても，部分デコードデータのみの Two-stream CNN を用いれば，同程度の認識率を達成できると考えられる．

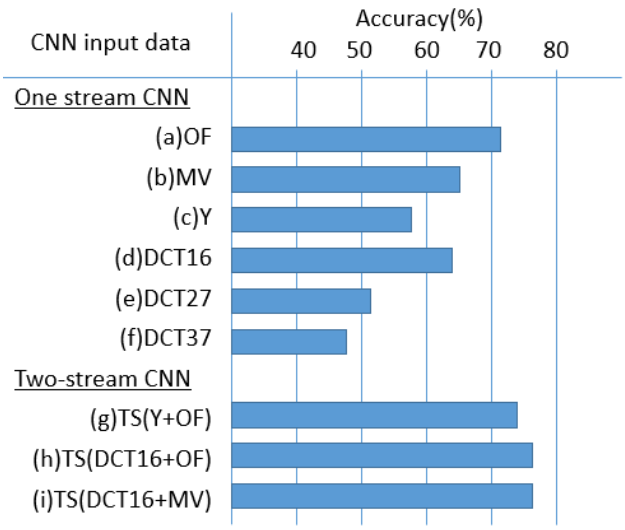

図 4　実験結果

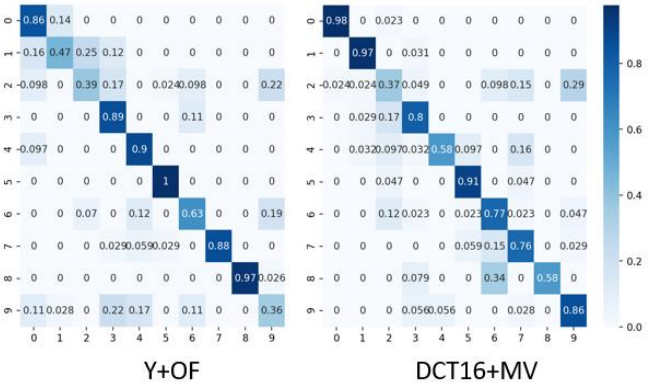

図 5　混同行列

4. おわりに

本検討では，映像符号化データに含まれる DCT 係数と動きベクトル情報を用いた Two-stream CNN を検討し，圧縮データであっても従来の手法と同等の認識率が得られることを明らかにした．今後は，使用する DCT の周波数成分の選択や，色差信号の DCT 係数の利用，様々な符号化モード情報の利用等について検討を進める予定である．

参考文献

[1] S. Karen and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," NIPS'14, pp. 568-576, 2014.
[2] 中村光貴，澤田友哉，杉本和夫，“時系列差分情報を付与した軽量な物体認識手法,”PCSJ/IMPS2019, P-2-15, pp.84-85, 2019.
[3] S. Khurram, A. R. Zamir , M. Shah, “UCF101:A dataset of 101 human actions classes from videos in the wild,” CoRR,abs/1212.0402, 2012.