VVC と CNN を組み合わせた YOLO-v7 のための映像符号化手法

Video Coding Scheme for YOLO-v7 Combining VVC and CNN

進藤嵩紘　渡部泰樹　渡辺裕
Takahiro Shindo　Taiju Watanabe　Hiroshi Watanabe

早稲田大学基幹理工学部
School of Fundamental Science and Engineering, Waseda University

1. まえがき

近年，画像認識技術の発達により，AI を用いた映像解析が急速に拡大している．そこで 2019 年，Moving Picture Experts Group (MPEG) では，Video Coding for Machines (VCM) を画像認識のための映像符号化と位置づけ，標準化作業を開始している．VCM では，より高い映像の圧縮率と画像認識精度が求められる．本稿では，CNN と Versatile Video Coding (VVC) を組み合わせることにより，YOLO-v7 による物体検出精度が高くなる映像符号化手法を提案する．YOLO-v7 の学習済みモデルにより，映像から抽出する特徴量を用いて CNN を学習することで，物体検出に有効な映像を作成する．

2. 提案手法

VVC[1]は最新の動画像符号化方式であり，高い圧縮率と映像品質を達成している．しかし，視聴用符号化方式として設計されているため，CNN と VVC を組み合わせることにより，YOLO-v7[2]による物体検出のための映像符号化手法を提案する．まず CNN により映像の画像サイズを半分に縮小し，その映像を VVC により圧縮することで高い圧縮率を達成する．VVC デコーダの出力映像を，CNN により元のサイズに戻すことで，物体検出に必要な映像情報の復元を目指す．VVC の符号化処理には VTM10.0[3]を使用し，参照構造は「lowdelay_P」とする．提案する映像符号化手法を図 1 に示す．

CNN の学習には YOLO-v7 の特徴量を用いる．映像の特徴抽出には学習済みモデルの backbone を使用し，生成映像と正解映像の特徴量の平均二乗誤差 (MSE) を損失計算に用いる．学習に用いる損失関数を次の式(1)に示す．

$$LOSS = MSE\big(yolo(I_{coded}), yolo(I_{raw})\big) \qquad (1)$$

ここに，$yolo$ は YOLO-v7 の backbone を用いた特徴抽出器を表し，I_{coded} は生成画像，I_{raw} は正解画像を表す．

3. 実験と結果

学習には, SJTU[4]，UVG[5]，MCML-4K-UHD[6]の三つのデータセットを用いる．テストでは VCM の Common Test Condition (CTC)[7] で用いられる，SFU-HW-Objects-v1[8]を使用する．学習に用いるシーケンスはすべて画像サイズが 4K であるため，テストでは最も画像サイズが大きい class A の Traffic シーケンスを用いる．

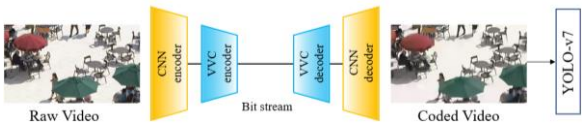

図 1　提案する映像符号化手法

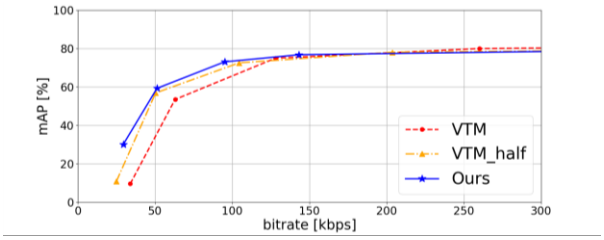

図 2　各符号化手法における bitrate と mAP の関係

提案手法による符号化映像と VTM10.0 による符号化映像の物体検出精度を比較する．また，画像サイズを半分に縮小し，VTM10.0 により符号化した場合の物体検出精度とも比較する．物体検出モデルは YOLO-v7 とし，比較する VTM の参照構造は「randomaccess」とする．検出精度は mean Average Precision (mAP) を用いて計測し，計測時に用いる Intersection over Union の閾値は 0.5 とする．符号化映像の bitrate と mAP の関係を図 2 に示す．図 2 より，提案手法は，映像の圧縮効率と物体検出精度で，VVC を上回ることが分かる．さらに，単純に画像サイズを縮小してから符号化する場合よりも，物体検出精度が高いことが分かる．

4. むすび

本稿では，YOLO-v7 のための映像符号化手法として，CNN と VVC を組み合わせた手法を提案した．CNN を YOLO-v7 の学習済みモデルを用いて学習させることで，物体検出に有効な映像を作成できることを実験により示した．

謝辞

本研究成果は，国立研究開発法人情報通信研究機構の委託研究 (05101) により得られたものである．

参考文献

[1] Versatile Video Coding, Standard ISO/IEC 23090-3, ISO/IEC JTC 1, Jul. 2020.
[2] C. Y. Wang, *et al.*, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for realtime object detectors.” arXiv preprint arXiv:2207.02696, 2022.
[3] S. K. J. Chen, *et al.*, Algorithm description for Versatile Video Coding and Test Model 10 (VTM 10). JVETS2002, 2020.
[4] L. Song, *et al.*, “The SJTU 4K Video Sequence Dataset,” in International Conference on Quality of Multimedia Experience, 2013.
[5] A. Mercat, *et al.*, “UVG dataset: 50/120fps 4K Sequences for Video Codec Analysis and Development,” in ACM Multimedia, 2020.
[6] M. Cheon, *et al.*, “Subjective and Objective Quality Assessment of Compressed 4K UHD Videos for Immersive Experience,” in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, 2018.
[7] Common test conditions for video coding for machines ISO/IEC JTC 1/SC 29/WG 04, Nov. 2022.
[8] H. Choi, *et al.* “A dataset of labelled objects on raw video sequences.” Data in Brief, 34:106701, 2021.