符号化雑音環境下における物体検出精度の改善手法
A Method for Improving Object Detection Accuracy in Coding Noise Environment

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

‡ 早稲田大学大学院基幹理工学研究科
‡Graduate School of Fundamental Science and Engineering, Waseda University

Abstract: Research and standardization activities for Video Coding for Machine (VCM) has been intensified. In this paper, we propose a method to improve the accuracy of image recognition by processing the coding noise in VVC encoded video. The proposed method is based on ESRGAN which is a Convolutional Neural Network (CNN). The evaluation method is the accuracy of object detection by YOLOv7. Experimental results show that the proposed coding noise processing improves object detection accuracy.

1 はじめに

近年，画像認識のための動画像符号化技術に関する研究が行われている．Versatile Video Coding(VVC)[1] による符号化映像には符号化雑音が加わるため，画像認識精度の低下を招く．本稿では，ニューラルネットワークを用いて符号化映像の符号化雑音を処理することにより，画像認識の精度を改善する手法について提案する．提案手法は，Enhanced Super-Resolution Generative Adversarial Networks(ESRGAN)[2] の生成器の構造をもとに作成する Convolutional Neural Network (CNN) である．評価手法には YOLOv7[3] の学習済みモデルによる物体検出精度を用いる．提案する符号化雑音処理により，物体検出精度が改善できることを実験により示す．

2 従来手法

Enhancing VVC Through CNN-Based Post-Processing [4] では，CNN を用いた雑音除去手法により，VVC による符号化映像の品質向上手法を提案する．ネットワーク構造は SRGAN[5] の生成器を参考に構成された CNN である．ネットワークは VVC による符号化映像のフレームを入力とし，出力画像と符号化前の画像との絶対誤差を用いて学習する．評価手法では PSNR(Peak Signal to Noise Ratio) を用い，CNN を用いた符号化雑音処理により，VVC による符号化雑音が低減することを示す．

3 提案手法

従来手法は，VVC による符号化映像の符号化雑音を低減し，符号化前の映像に近づけることを目的とする．しかし，低減できる符号化雑音の大きさは小さく，画像認識精度の改善には結びつかない．そこで，VGG[6] により得られる特徴量を損失計算に用いることで，VVC による符号化映像の画像認識精度の改善を目指す．また，ESRGAN の生成器の構造を参考に，SRGAN を参考に構成される従来手法のネットワーク構造を改変する．

3.1 モデル構造

ESRGAN は SRGAN と同じく，画像の超解像を行うモデルである．このモデルは，SRGAN の residual block[7] を Residual-in-Residual Dense Block(RRDB) に置き換えた構造を持ち，より深いニューラルネットワークを構成することで，画像の超解像において細かな絵柄の再現を可能にする．本稿の提案手法でも，従来手法の residual block を RRDB に置き換たモデルを用いる．提案するネットワーク構造を図 1 に示す．

3.2 損失関数

VGG は画像認識精度の高さから，特徴抽出手法として広く利用されるモデルの一つである．出力画像と符号化前の画像の絶対誤差を損失計算で用いる従来手法とは異なり，提案手法ではそれらの平均二乗誤差と，それらから VGG により抽出される特徴量の平均二乗誤差を損失計算に用いる．従来手法で用いる損失を式 (1) に，提案手法で

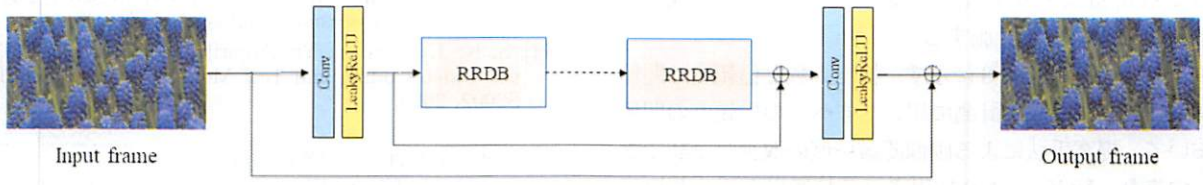

図 1: 提案手法のモデル構造

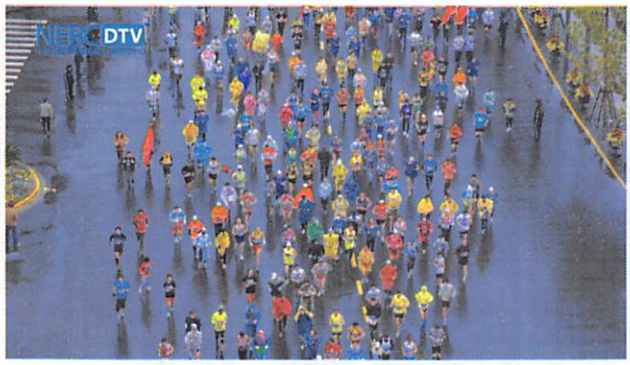

図 2: Marathon シーケンス

(a)正解フレーム (b)入力フレーム (c)出力フレーム

図 3: 検出結果

用いる損失を式 (2) に示す.

$$loss = l_{L1} \tag{1}$$

$$loss = l_{MSE} + l_{VGG} \tag{2}$$

4 実験及び結果

提案手法により，符号化雑音環境下において画像認識精度が改善することを実験により示す．学習用データセットには，SJTU データセット (Marathon シーケンスは除く)[8]，UVG データセット [9]，MCL-JCV データセット [10] を用いた．これらのデータセットを VTM10.0[11] を用いて符号化する．参照構造はすべてランダムアクセス，量子化係数は 37 である．符号化後の映像をモデルの入力とし，VGG の学習済みモデルにより抽出される特徴量を用いてモデルの学習を行う．テスト用データセットには，SJTU データセット内の Marathon シーケンスを用いる．1 秒間のフレーム数は 25 であり，総フレーム数は 60 である．Marathon シーケンスのフレーム画像を図 2 に示す．テストデータの符号化方法は学習に用いるデータの符号化手法と同一である．

評価手法は入力動画像と出力動画像と正解動画像のそれぞれに対して，YOLOv7 の学習済みモデルにより人物を検出した結果を用いる．検出時に用いる信頼度の閾値を 0.25 から 0.95 まで 0.05 刻みで変化させたときに，検出される人数をそれぞれ計測する．

検出結果の一例を図 3 に示す．検出人数と信頼度の閾値の関係を図 4 に示す．計測に用いたすべての信頼度の閾値において，提案手法による画像認識精度の改善が確認できる．つまり，VGG により抽出される特徴量を用いた符号化雑音処理は，符号化映像の画像認識精度の改善に有効である．

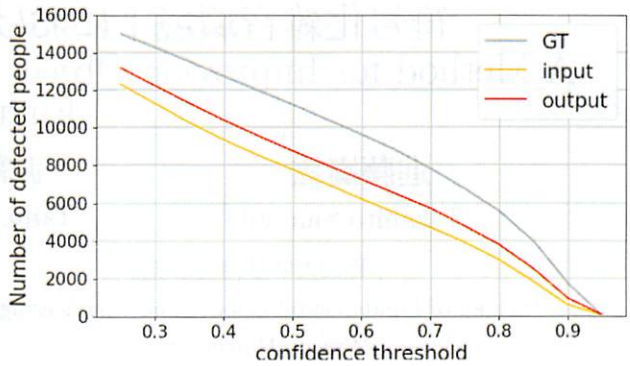

図 4: 検出人数と信頼度の閾値の関係

5 まとめ

VGG による特徴抽出を用いた符号化雑音処理により，符号化映像の画像認識精度を改善できることを示した．VVC による符号化映像を提案手法により処理することで，YOLOv7 による人物検出精度が改善することを実験により確認した．今後，提案手法の汎用性を確かめるために，テスト用のシーケンスを増やす必要がある．

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] S. L. B. Bross, J. Chen, Versatile Video Coding (Draft 10). JVET-S2001, 2020.

[2] X. Wang *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks", ECCV Workshop, 2018.

[3] Chien-Yao Wang *et al.*, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for realtime object detectors" arXiv preprint arXiv:2207.02696 (2022).

[4] F. Zhang *et al.*, "Enhancing VVC Through CNN-Based Post-Processing", ICME, 2020.

[5] C. Ledig *et al.*, "Photo-realistic single image superresolution using a generative adversarial network", CVPR, 2017.

[6] K. Simonyan *et al.*, "Very deep convolutional networks for large-scale image recognition", ICLR, 2015.

[7] K. He *et al.*, "Deep residual learning for image recognition", CVPR, 2016.

[8] L. Song *et al.*, "The SJTU 4K video sequence dataset", QoMEX, 2013.

[9] A. Mercat *et al.*, "UVG dataset: 50/120fps 4K sequences for video codec analysis and development", ACM Multimedia Systems Conference, 2020.

[10] H. Wang *et al.*, "MCL-JCV: A JND-based H.264/AVC video quality assessment dataset", ICIP, 2016.

[11] S. K. J. Chen, Y. Ye, Algorithm description for Versatile Video Coding and Test Model 10 (VTM 10). JVET-S2002, 2020.

早稲田大学基幹理工学部情報通信学科
〒*162-0072* 東京都新宿区大久保 *3-14-9* 早大 *66-401*
Phone: 03-5286-2509, E-mail: taka_s0265@ruri.waseda.jp