空間と時間的一貫性のある動画表現の一検討

A Study of Spatially and Temporally Consistent Video Representation

速見 泰雅† 進藤 嵩紘† 渡辺 裕†

Taiga Hayami† Takahiro Shindo† Hiroshi Watanabe†

† 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: Implicit Neural Representation（INR）は、信号をネットワークに埋め込むことでその信号を表現する手法である。特に、動画を埋め込んだネットワークのモデル圧縮は動画圧縮として機能し、動画圧縮手法としても活用できる。従来のフレーム時刻やフレームから抽出される特徴量を用いた動画の INR では、空間情報や時間情報がそれぞれ十分に考慮されていないため、再構成された動画において一貫性が失われることがある。本稿では、再構成動画の空間・時間的一貫性を維持するために、フレームの高周波成分と隣接フレーム情報を活用したネットワークによる動画表現手法を提案する。

1 はじめに

近年、インターネットとそれに伴う動画配信サービスの普及により、高品質な動画の需要が高まっている。特に動画は画像や音声に比べ情報量が多く、高品質な動画を扱うためには効率的な動画圧縮技術が不可欠である。Implicit Neural Representation (INR) は、従来の HEVC[1]/VVC[2] といったルールベースの動画圧縮技術や、深層学習を用いた学習ベースの動画圧縮手法とは異なるアプローチにより動画圧縮を実現でき、注目されている。INR は様々な信号をネットワークに埋め込むことで表現する手法である。特に動画を扱う Implicit Neural Representation for Videos (INRV) においては、埋め込まれたネットワークをモデル圧縮することで動画圧縮を実現することができる。従来の INRV 手法は効率的なネットワークへの埋め込み方法に焦点を当てており、フレームの時刻やフレームから抽出される特徴量をネットワークの入力に用いる。そして、各入力情報に対応するフレームをネットワークの出力として再構成する。しかし、これらの手法は空間情報や時間情報をそれぞれ十分に考慮していないため、再構成された動画において一貫性が失われることがある。

本稿では、これらの一貫性を保つために、フレームの高周波成分と隣接フレーム情報を活用する映像表現手法を提案する。実験では、従来手法と比べて再構成された動画の品質と圧縮性能が改善されることが確認された。

2 関連研究

2.1 Implicit Neural Representation

INR は複雑な信号を簡素なネットワークに埋め込むことで表現できることから、明示的な表現の難しい 3D Shape や Scene へ適用されてきた。これらはインデックスベース手法を用いており、各座標から対応する色や密度をネットワークにより推定する。動画を扱う INRV において、インデックスベース手法はフレームサイズが大きくなると入力ペアが莫大に増加するため、フレームベース手法が一般に研究されている。この手法は各フレームの情報からそのフレームをネットワークにより推定する。フレーム情報にはフレーム時刻を表すタイムスタンプ [3] やフレームからエンコーダを用いて抽出された特徴量 [4] が使用される。後者は、タイムスタンプを使用する手法と比べ、ネットワークの入力が各動画に依存するため、より精巧に動画を再構成することができる。

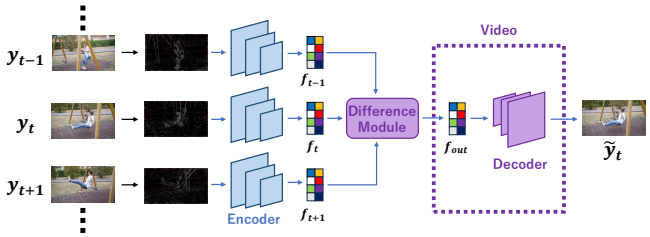

図 1: 提案手法の構造

2.2 Video Compression

従来のルールベースの動画圧縮技術や深層学習による学習ベースの動画圧縮手法は、近年圧縮効率が向上している一方、構造が複雑化し、デコード時の速度やコストが課題となっている。INRV は単純なネットワークを使用しているため、従来の動画圧縮手法と比べてデコード速度が速く、コストが低い。従来の INRV 手法は動画の効率的なネットワークへの埋め込み方法に焦点を当てている。動画圧縮においては、動画が埋め込まれたネットワークパラメタに量子化や枝刈、エントロピー符号化を適用することで実現する。

3 提案手法

再構成動画における空間的・時間的な一貫性を維持するために、フレームの高周波成分から得られる特徴量と、隣接フレーム情報を用いた映像表現手法を提案する。提案手法の構造を図 1 に示す。特徴量ベースの方法では効率よく特徴量を抽出する必要がある。フレーム全体から抽出する手法では、連続するフレームの類似性から、それぞれ抽出される特徴量も類似していることが推察され、冗長であ

表 1: 再構成動画の品質評価

Method	PSNR ↑	MS-SSIM ↑	LPIPS ↓
NeRV(All)[3]	28.91	0.8865	0.3663
HNeRV(All)[4]	30.94	0.9160	0.2955
HNeRV(Suc)[4]	31.28	0.9236	0.2910
Ours(All)	31.50	0.9267	0.2874
Ours(Suc)	31.45	0.9265	0.2871

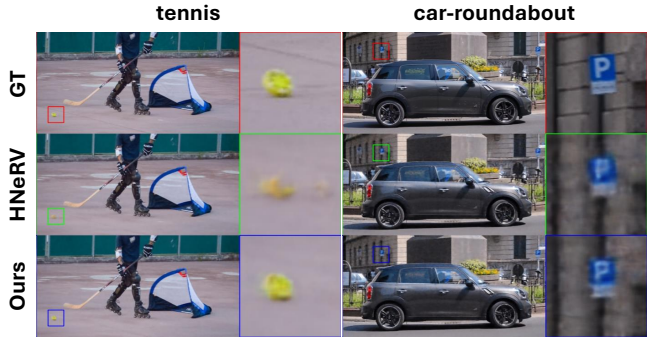

図 2: 再構成フレームの可視化例

る。そのため、類似した特徴量から類似性の低いフレームの高周波成分を再構成することは困難となる。また、深層学習における Spectral Bias [5] によりネットワークはフレームの高周波成分を学習しにくい傾向にある。そこで、我々はフレームの高周波成分から特徴量を抽出する。高周波成分は連続するフレームにおいて類似性が低く、特徴量の冗長性を削減することができる。また、特徴量ベース手法は時間情報を明示的にネットワークに与えないため、フレーム関係を学習するのが難しく、再構成動画の時間的な一貫性が失われる場合がある。そこで、隣接フレームから同様に抽出された特徴量を図 1 の Difference Module で活用する。対象フレームの特徴量 f_t と隣接フレームから得られる特徴量 f_{t-1}、f_{t+1} は、

$$
\begin{aligned}
f_{diff} &= |f_{t+1} - f_{t-1}|, \\
f_{out} &= f_t \cdot \{\frac{f_{diff}}{\max(f_{diff})} \times s + (1-s)\},
\end{aligned}
\tag{1}
$$

により融合され f_{out} が得られる。なお、s はハイパーパラメタで $s = 0.1$ とする。これにより、再構成動画の時間的な一貫性を保持する。

4 実験

評価には DAVIS Dataset [6] を用いる。このデータセットはフレーム枚数 25-104 枚からなる 50 本の動画で、フレームサイズは 1080×1920 である。実験では 640×1280 に切り取って使用する。ハイパスフィルタではフレームの 80% の低周波成分が除去される。学習は 300 エポック行い、Adam オプティマイザーにより最適化される。損失は

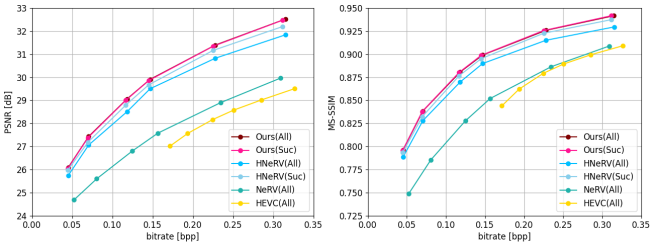

図 3: 動画圧縮結果

再構成フレームと正解画像との L2 損失を用いる。モデル圧縮では、特徴量とデコーダのパラメタを量子化をし、ハフマン符号化を適用する。特徴量とデコーダパラメタの量子化係数はそれぞれ 6、8 とする。比較手法としてタイムスタンプを使用する NeRV[3] とフレーム全体から抽出される特徴量を使用する HNeRV [4] を用いる。評価指標は PSNR、MS-SSIM、LPIPS を用いる。

再構成動画 50 本の平均品質評価結果を表 1 に示す。HNeRV ではその中の 1 本の動画で学習が十分に収束しなかった。学習が上手くいく 49 本の平均結果を追記する (Suc)。また、再構成フレームの可視化例を図 2 に示す。提案手法が従来手法に比べ詳細な部分を再構成できていることが確認できる。図 3 にモデルサイズを変更し、動画圧縮性能を比較した結果を示す。ルールベースの HEVC に比べ圧縮性能が優れていることが確認できる。

5 まとめ

再構成動画の空間・時間的一貫性を維持するために、フレームの高周波成分と隣接フレーム情報を用いた動画表現を提案し、その有効性を示した。提案手法にはいくつかのハイパーパラメタがあり、各動画に適切な値を設定する手法を検討する必要がある。

参考文献

[1] High Efficiency Video Coding, Standard ISO/IEC 23008-2, ISO/IEC JTC 1, Apr.2013.

[2] Versatile Video Coding, Standard ISO/IEC 23090-3, ISO/IEC JTC 1, Jul. 2020.

[3] H. Chen *et al.*, "NeRV: Neural Representations for Videos," Advances in Neural Information Processing Systems, 21557-21568, Nov. 2021.

[4] H. Chen *et al.*, "HNeRV: A Hybrid Neural Representation for Videos," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10270-10279, Jun. 2023.

[5] N. Rahaman *et al.*, "On the Spectral Bias of Neural Networks," Proceedings of Machine Learning Research, 5301-5310, Jun. 2019.

[6] F. Parazzi *et al.*, "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 724-732, Jun. 2016.

早稲田大学 基幹理工学研究科 渡辺研究室
〒*169–0072* 東京都新宿区大久保 *3-14-9*
Phone: 03-5285-2509, Fax: 03-5286-3488
E-mail: hayatai17@fuji.waseda.jp