画素値の動的変化に基づく動画像表現
Video Representation Based on Dynamic Shifts in Pixel Values

進藤嵩紘 † 速見泰雅 † 田中頌子 † 渡辺裕 †
Takahiro Shindo† Taiga Hayami† Shoko Tanaka† Hiroshi Watanabe†
† 早稲田大学
†Waseda University

Abstract: Implicit Neural Representations (INRs), a signal coding method, has been gaining popularity. This method embeds signals within a model to represent them. This paper explores a method of representing videos using INRs. We construct a model that takes coordinates as inputs and outputs the corresponding RGB signals for the video.

1. はじめに

近年，動画配信サービスや SNS などの利用拡大に伴い，高解像度・高フレームレートの動画を視聴する機会が増加している．このような動画利用の需要に応えるためには，映像の符号化技術が不可欠である．そこで MPEG により，HEVC/H.265 [1] や VVC/H.266 [2] などのルールベースの映像符号化アルゴリズムが標準化されてきた．また，ニューラルネットワークを用いた動画像符号化手法の研究も盛んである．その手法の一つに，Implicit Neural Representations (INRs) を用いた映像表現手法，Neural Representations for Videos (NeRV) [3] が存在する．NeRV は時間情報 t を入力として，その時間に対応する動画中のフレームを出力するニューラルネットワークにより，動画を表現する．一方，INRs を用いた静止画像表現手法である Compression with Implicit Neural Representations (COIN) [4] は，NeRV とは異なる入力と出力の特徴をもつ．COIN は，入力として x 座標と y 座標をもち，それに対応する画像内の RGB 値を出力する．

そこで本稿では，NeRV と COIN を組み合わせた，映像の表現手法について検討する．x 座標と y 座標を入力として，それに対応する動画像の RGB 信号を出力するニューラルネットワークを作成する．実験において，提案手法を NeRV と比較し，その有効性について検証する．

2. 従来手法

2.1. Implicit Neural Representations (INRs)

INRs は，音声，画像，映像，3D shape などの様々な信号を符号化する手法として，ここ数年でますます注目されている．これは，信号をニューラルネットワークに埋め込むことで，それらを表現する手法である．ネットワークそのものが埋め込まれた信号情報であるため，モデルのサイズを縮小することは，信号のサイズの縮小を意味する．そこで，ニューラルネットワークの量子化や枝刈りを用いて，埋め込まれた信号の圧縮することが可能である．INRs を用いた動画の圧縮手法は，ルールベースの手法に匹敵する性能を持つことが多くの研究により示されている．

2.2. Neural Representations for Images/Videos

INRs を用いた画像と映像の表現手法として，それぞれ COIN と NeRV が提案されている．COIN は，座標情報を入力として，ある特定の静止画像を復号するモデルを使用する．ネットワークの入力と出力の関係は，次の式で表される．

$$f_\theta(x, y) = (r, g, b). \tag{1}$$

式 (1) において，x, y は入力座標情報，r, g, b はその座標に対応する画素値を表す．θ は COIN のモデルの重みを表し，f_θ は重み θ を使用した，モデルの関数を表す．

NeRV は，時刻情報を入力として，ある特定の映像を復号するモデルにより，映像を表現する．モデルの入力と出力の関係は，次の式で表される．

$$f_\theta(t) = v_t. \tag{2}$$

式 (2) において，t は入力時刻情報，v_t はその時刻に対応するフレームを表す．θ は NeRV のモデルの重みを表し，f_θ は重み θ を使用した，ネットワークの関数を表す．

NeRV と COIN では，再構成した信号が元々の信号に等しくなるように，ネットワークを学習させる．つまり，出力と，埋め込み対象となる信号の誤差を小さくするように，重み θ を更新する．

3. 提案手法

私たちは，x，y 座標を入力とする，映像表現のためのモデルを提案する．これは，COIN と同様の入力情報をもつ一方で，NeRV と同様に映像表現を行うモデルである．モデルは CNN を用いて構成する．提案するモデルにおける，入力と出力の関係を図 1 に示す．また，この関係は次の式で表される．

$$f_\theta(x, y) = (\boldsymbol{r}, \boldsymbol{g}, \boldsymbol{b}). \tag{3}$$

式 (3) において，x, y は入力座標情報，$\boldsymbol{r}, \boldsymbol{g}, \boldsymbol{b}$ はその座標に対応する全ての映像フレームの画素値を表す．例えば、出力 $\boldsymbol{r}$ は，フレーム数 T の映像の場合，$\boldsymbol{r} = \{r_1, r_2, ...r_T\}$ のように表される．r_i は，i 番目のフレームにおける赤色

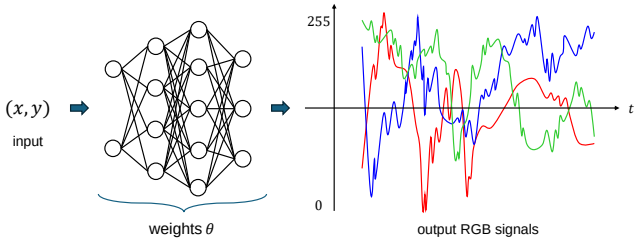

Fig. 1. 提案手法におけるモデルの入力と出力の関係.

Table 1. 提案手法と従来手法における，パラメータ数と再構成画像の品質 (PSNR[dB]) の関係.

method	param[M] (↓)	PSNR[dB] (↑) Bosphorus	ReadySetGo
NeRV	2.69	32.96	25.44
Ours	**2.58** (-4.10[%])	**36.33** (+3.37)	**28.14** (+2.70)

信号の画素値である．θ を提案するモデルの重みとし，f_θ を重み θ を使用したネットワークの関数とする．このモデルは，ある座標情報を受け取ったとき，その座標に対応するすべてのフレームの RGB 値を出力する．パラメータ θ の更新には，復号信号と正解信号の二乗誤差を使用し，対象となる RGB 値を復号するように学習させる．

4. 実験及び結果

提案手法による動画像の表現性能を評価する．UVG データセット [5] における，Bosphorus シーケンスと ReadySetGo シーケンスを評価に用いる．これらのシーケンスを，高さ 270[pixel]，幅 480[pixel] にリサイズする．また，フレームレートが 24[fps] となるように，フレーム数を削減する．削減後のフレーム数は 120 フレーム，5 秒間の動画像である．この様に用意した二種類のシーケンスを，提案手法と従来手法 (NeRV) により表現する．NeRV の公式実装を比較手法として使用する．提案手法のモデルは，可能な限り NeRV のモデルサイズに近くなるように設計する．

提案手法と比較手法における，モデルのサイズと画像表現の性能の関係を表 1 に示す．使用する NeRV のパラメータ数は 2.69[M] であり，作成した提案モデルのパラメータ数は 2.58[M] である．これより，各モデルのサイズはほとんど等しいことが確認できる．また，表 1 より，再構成フレーム画像の品質においては，提案モデルの方が優れている．再構成画像の一例を図 2 に示す．図 2 より，提案モデルは NeRV よりも，色や物体の形の表現力に長けていることが分かる．これらの結果より提案モデルは映像表現性能において優れていることが確認できる．NeRV と比較して，提案モデルが，より効率よくネットワークのパラメータを利用していることが分かる．

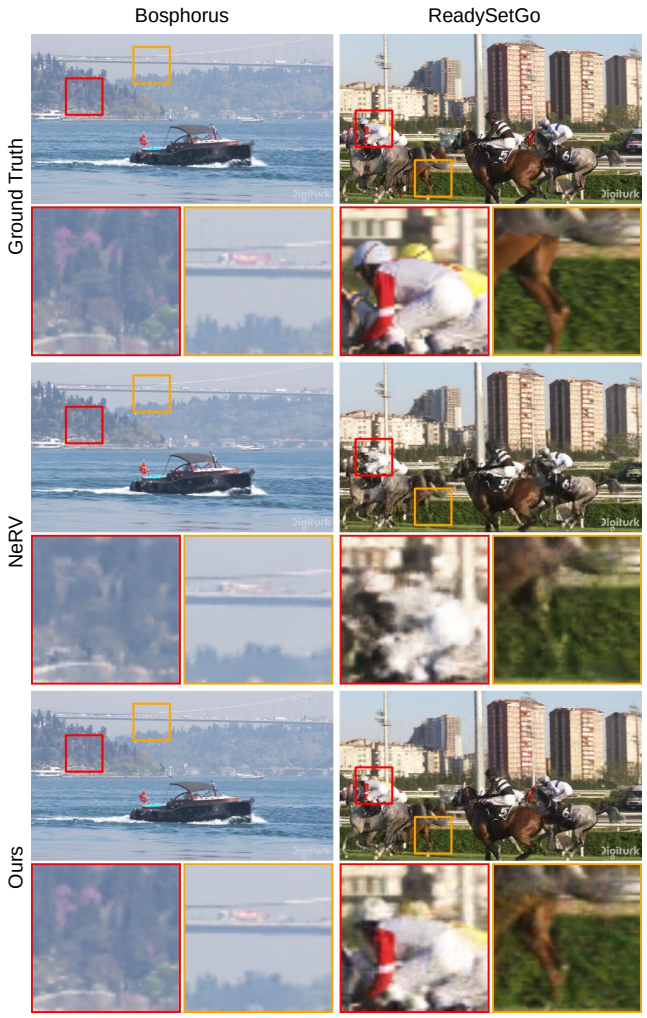

Fig. 2. 各シーケンスにおける再構成フレーム画像.

5. まとめ

本稿では，座標情報を入力とした動画像表現モデルを作成し，その精度を実験により測定した．NeRV と提案手法を比較することにより，本手法の有効性を確認した．今後の研究では，テストシーケンス数を増やした上で，モデル構造について見直し，より最適なモデルの設計方法について検討する必要がある．

6. REFERENCES

[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] C. Hao *et al.*, "NeRV: Neural Representations for Videos," Advances in Neural Information Processing Systems, 34, 2021, 21557-21568.

[4] D. Emilien *et al.*, "Coin: Compression with implicit neural representations," arXiv preprint arXiv:2103.03123, 2021.

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

早稲田大学大学院 基幹理工学研究科 情報理工・情報通信専攻
〒*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