1U-07

ビデオ伝送に対する深層学習の適用可能性に関する検討

渡邊 翔太 † 川崎 慈英 ‡ 猿渡 俊介 ‡ 渡辺 尚 ‡
† 大阪大学工学部 ‡ 大阪大学大学院情報科学研究科

1 はじめに

スマートフォンの普及によりコミュニケーションの手段としてビデオ通話が頻繁に使われるようになった．一般的に，映像データはデータ量が大きいため，ビデオ通話の使用率の上昇は通信帯域のひっ迫を招く．本稿では，ビデオ通話の映像トラヒックを削減する方式として，深層学習によるモザイク復元技術を用いた手法を提案する．提案システムは送信データサイズを 90%以上削減しつつ，SSIM による評価で 10,000 回の学習で約 85%の復元率を達成した．

2 システムモデル

図 1 に本研究で対象とする映像伝送のモデルを示す．送信側のカメラから受信側のディスプレイに対してネットワークを介して映像を伝送するモデルを考える．この時，送信側では符号化，受信側では復号化を行うことでネットワーク上に伝送されるデータ量を減らすことができる．

現在の映像伝送では，H.264/AVC や H.265/HEVC などの符号化技術を用いて圧縮を行っている．本稿では，これら従来の画像圧縮方式と共存してさらにデータ量を削減する方法を模索する．

3 提案手法

提案システムの基本的なアイディアは，送信側が映像を低解像度化して映像を伝送することで通信量を減らし，受信側が深層学習を用いて低解像度の映像から元の高解像度の映像を復元することである．ビデオ通話に特有の「同じ人の顔画像が連続して送信される」という特性を利用する．具体的には，同じ人の顔画像のみを学習したモデルを用いることで，モザイク復元における様々な復元パターンを減らして精度向上や学習の効率化を測る．モザイク画像自体は既存の画像圧縮を用いて伝送できるため，本提案手法は既存の画像圧縮と共存できる．

モザイク復元には DCGAN (Deep Convolutional Generative Adversarial Network) [1] を用いる．DCGAN は生成器と判別器の 2 つのネットワークを用いた生成モデルである GAN (Generative Adversarial Network)[2] に CNN (Convolutional Nural Network) を利用した画像生成モデルである．判別器は与えられたデータが生成器によって作られた偽物のデータか入力された正解データかを判別する．生成器は判別器が入力データと誤認させるような入力データとよく似たデータを生成する．これらの 2 つのネットワークが交互に精度を高め合うことによって，最終的に生成器は入力データと非常によく似たデータを生成することができる．

本システムは，学習通信フェーズと高解像度化通信フェーズの 2 つのフェーズから構成される．学習通信フェーズにおける受信器の学習が終了すると，受信器は学習が終了したことを知らせる信号を送信器に送ることにより高解像度化通信フェーズに移行する．

図 2 に学習通信フェーズを示す．学習通信フェーズでは，受信器は高解像度の映像のディスプレイ表示と学習モデルの生成を同時に行う．送信器は高解像度の映像をそのまま受信器に対して送信する．高解像度の映像を受け取った受信器は，受信した高解像度の映像をそのままディスプレイに表示する．受信器はある一定時間の映像を低解像度画像に変換し，低解像度画像と変換前の高解像度画像の対を用いて学習モデルを生成する．

図 3 に高解像度化通信フェーズを示す．高解像度化通信フェーズでは，送信器は高解像度の映像を低解像度化したものを受信器に対して送信する．受信器は受け取った低解像度の映像を学習通信フェーズで生成した学習モデルを利用して高解像度画像へ復元する．学習モデルを利用して復元した復元画像はディスプレイで表示される．

次に学習モデルを示す．学習モデルは図 2 の学習通信フェーズで生成されて図 3 の高解像度通信フェーズで利用される．まず，生成器のモデルを示す．1 層のバッチ正規化層と 2 層の ReLU 層，3 層の畳み込み層，4 層の結合層を 8 層まで繰り返し，9 層に 2 倍の Up Scale 層，10 層にバッチ正規化層，11 層に ReLU 層，12 層に転置畳み込み層がある．この 1 層から 12 層までをもう一度繰り返して 13 層から 24 層とする．さらに，25 層の畳み込み層と 26 層に ReLU 層をもう一度繰り返して最後の 29 層にシグモイド層がある．次に判別器のモデルを示す．1 層の畳み込み層と 2 層のバッチ正規化層と 3 層の ReLU 層の 3 つの層を 18 層まで繰り返し，19 層に畳み込み層，20 層に平均を取る層がある．

4 初期的評価

提案システムの基礎的な性能を確認することを目的として画像の低解像度化及び復元を行い，復元画像を低解像度画像と比較した．

4.1 評価環境

本評価に用いたテスト画像列は，PENTAX KS-2 を用いて撮影した 10 分間の動画から取得した．撮影した動画を ffmpeg を用いてフレームレート 30 の jpeg 画像に変換した．各 jpeg 画像の顔部分を Pillow を用いて抽出することで解像度 80×80 と 160×160 の 2 種類の入力画像群を作成した．解像度 80×80 と 160×160 の画像のサイズはそれぞれ 80×80 と 160×160 である．入力画像群は 200,000 枚作成したものの中から無作為に 16 枚を選択して学習に用いた．

画像の低解像度化及び復元には，srez[3] を用いた．srez の実行には OS は Ubuntu 16.04.1，CPU はインテル Xeon プロセッサー E5-2637 v3 を使用した．GPU は使用していない．srez では，入力したテスト画像を元にそれぞれの解像度を 1/4 にした解像度 20×20 と 40×40 の低解像度画像を作成，DCGAN を利用して画像を復元している．学習係数の初期値は 0.0002 を用いている．重みの初期値は正規分布から取得したランダム値，バイアスの初期値は 0 を用いている．学習パラメータの更新には Adam [4] を用いた．Adam の初期値は論文 [4] の奨励値を用いている．入力画像の内，学習に用いていない画像の中からランダムで 8 枚を評価に用いた．

4.2 主観評価

提案システムによって画像が復元できているかどうかを確認するために，高解像度画像，低解像度画像，復元画像を比較する主観評価を行った．図 4 に主観評価に用いたテスト画像を示す．図 4 の上段は解像度 80×80 の画像，下段は解像度 160×160 の画像の主観評価である．それぞれ左側にはテスト画像，中央にはモザイク処理をした画像，右側には復元画像を解像度とデータサイズと共に示している．学習は 20,000 回行った．図 4 から，復元画像は元画像と同じ人物を復元できていることが分かる．また，表情に関しては若干差があることも分かる．

4.3 定量評価

提案システムの基本的な性能を確認するために，データ量の削減量，画像の復元量，計算量の 3 つについて定量的に評価した．まず，提案システムによりデータ量がどの程度下げられたのか確かめるために，図 4

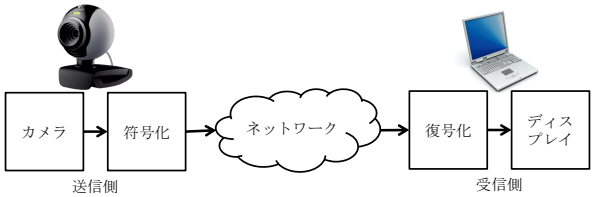

図 1: システムモデル

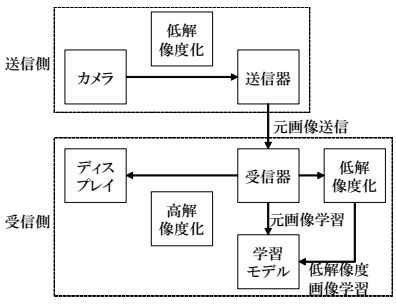

図 2: 学習通信フェーズ

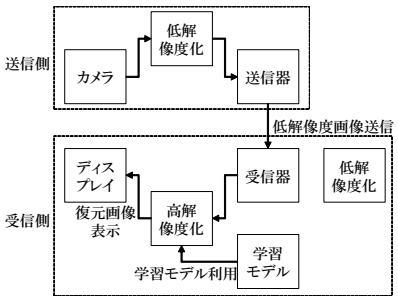

図 3: 高解像度化通信フェーズ

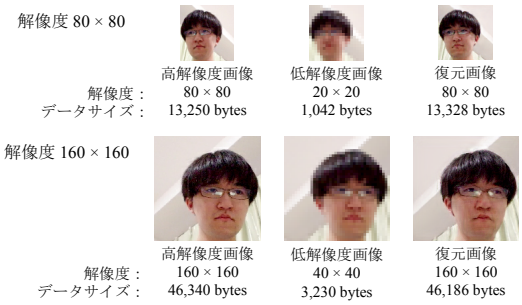

図 4: 主観評価に用いたテスト画像

Copyright ©2018 Information Processing Society of Japan.
All Rights Reserved.

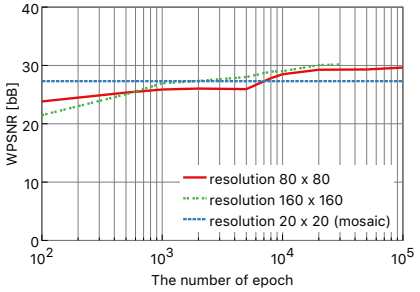

図 5: 学習回数と WPSNR との関係

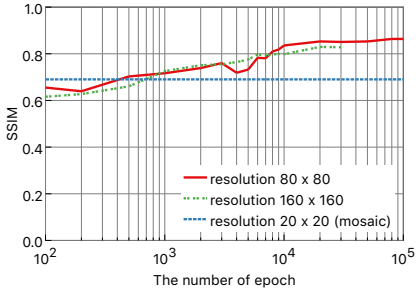

図 6: 学習回数と SSIM との関係

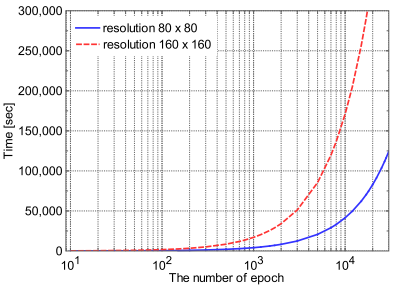

図 7: 学習回数と計算量の関係

の画像のデータサイズを調べた．図 4 における元画像と低解像度画像のデータサイズは解像度 80×80 の画像でそれぞれ元画像が 13,250 bytes，低解像度画像が 1,042 bytes であり，解像度 160×160 の画像でそれぞれ元画像が 46,340 bytes，低解像度画像が 3,230 bytes であった．どちらの解像度においても画像におけるデータ量を 90 %以上削減できている．また，この学習における学習モデルのデータサイズは約 92 MB であった．

次に，低解像度画像と復元画像の WPSNR (Weighted Peak Signal-to-Noise Ratio) を比較した．WPSNR とは，YCbCr 色空間 (Y:輝度, Cb:青色系統, Cr:赤色系統) のそれぞれの PSNR の真値に Y : Cb : Cr = 8 : 1 : 1 の重み付き平均を行ってデシベル値に直したものである [5]．YCbCr 色空間では Y が Cb, Cr よりも多くの情報を所持しているため，通常の PSNR ではなく WPSNR を用いて評価を行った．WPSNR を求めるために，以下の式を用いた．ただし，(n, m) は画像の縦横のピクセル幅，original(i, j) は高解像度画像におけるピクセル (i, j) の階調値，encoded(i, j) は復元後画像におけるピクセル (i, j) の階調値を表す．

$$\mathrm{PSNR} = -10 \log_{10} \frac{1}{nm} \sum_{i=0}^{n} \sum_{j=0}^{m} \frac{\mathrm{MSE}}{255^2}$$

$$\mathrm{MSE} = \{\mathrm{original}(i, j) - \mathrm{encoded}(i, j)\}^2$$

WPSNR は復元画像 8 枚それぞれについて PSNR を計算して，真値で平均を計算してからデシベル値とした．

図 5 に評価結果を示す．図 5 から次のことがわかる．1 つ目は，学習回数が増えるごとに解像度 80×80 の復元画像の元画像に対する WPSNR と解像度 160×160 の復元画像の元画像に対する WPSNR が向上していることである．学習回数が画像の復元量に影響することが考えられる．2 つ目は，復元画像の元画像に対する WPSNR が低解像度画像の元画像に対する WPSNR よりも高い事である．

次に，SSIM (Structual SIMilarity) [6] を用いて評価を行った．同じ位置の各画素の輝度値がどの程度変わったかを示す PSNR と比較して，輝度・コントラスト・構造を軸として各画素およびその周囲との相関を考慮した SSIM の方が人間の視覚的特性を反映できることが確認されている [7]．低解像度画像と復元画像の SSIM を比較するために，以下の式を用いた．

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

ここで，x, y はそれぞれ元の画像と符号化後の画像における各画素を要素とするベクトル，μ_x, μ_y はそれぞれ画像 x, y の平均画素値，σ_x, σ_y はそれぞれ画像 x, y の画素値の標準偏差，σ_{xy} は画像 x, y の共分散を表す．また，$C_1 = (255K_1)^2, C_2 = (255K_2)^2$ と表され，パラメータ K_1, K_2 は文献 [6] と同じ値 ($K_1 = 0.01, K_2 = 0.03$) を用いている．同様に文献 [6] と同じく評価前の画像にガウジアンフィルタをかけて前処理をしている．SSIM は 0 から 1 までの値をとり，全く同じ画像の時に 1 を示す．

図 6 に評価結果を示す．図 6 から WPSNR と同様に学習回数が増えるに従って SSIM が増加していることが分かる．

次に，学習通信フェーズと高解像度通信フェーズのそれぞれにおける計算量について評価を行った．図 7 に学習通信フェーズにおける評価結果を示す．図 7 では，解像度 80×80 の画像と解像度 160×160 の画像それぞれの学習回数に対する計算時間を示している．図 7 から次のことがわかる．学習回数が 10,000 回程度を達成するのに，解像度 80×80 の画像では 50,000 秒 (約 13 時間)，解像度 160×160 の画像では 175,000 秒 (約 48 時間) かかる．解像度 (画像サイズ) を大きくすると計算に時間がかかるため，画像サイズの小さい画像を利用する必要があると考えられる．尚，高解像度通信フェーズでは，画像 16 枚を生成する計算時間が約 1.14 秒であった．

5 関連研究

本研究は単一画像の超解像技術と深層学習の生成モデルに関連する．単一画像の超解像技術としては，A+ (Adjusted Anchored Neighborhood Regression for Fast Super-Resolution)[8]，SRCNN (Image Super-Resolution Using Deep Convolutional Networks)[9]，RAISR (Rapid and Accurate Image Super Resolution)[10] が研究されている．これらの研究では解像度を縦横各 2 倍の計 4 倍にする試みが行われている．それに対して本研究では，DCGAN を用いてより低解像度の画像から高解像度の画像を復元することを目指している．例えば，4 節の評価では縦横各 4 倍の計 16 倍の高解像度化を試みている．

6 おわりに

本稿では，ビデオ通話の映像トラヒックを削減する方式として，深層学習によるモザイク復元技術を用いた方式を提案した．提案システムにより低解像度画像を正しく復元することができた．送信データサイズを 90%以上削減しつつ，SSIM による評価で 10,000 回の学習で 85%程度の復元率を達成した．

謝辞

本研究は JSPS 科研費 (JP16H01718，JP17J02859，JP17KT0042，JP16K16044)，NTT アクセスサービスシステム研究所の支援の下で行った．

参考文献

[1] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in *Proceedings of International Conference on Learning Representations 2016 (ICLR’16)*, 2016.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proceedings of Advances in Neural Information Processing Systems 27 (NIPS’14)*, pp. 2672–2680, Curran Associates, Inc., 2014.

[3] D. Garcia, “Image super-resolution through deep learning.” https://github.com/david-gpu/srez, 2016. Access: 2017/12/21.

[4] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of The International Conference on Learning Representations 2015 (ICLR’15)*, 2015.

[5] Z. Wang, L. Lu, and A. C. Bovik, “Video quality assessment based on structural distortion measurement,” *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, 2004.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[7] Z. Wang and A. C. Bovik, “Mean squared error: Love it or leave it? a new look at signal fidelity measures,” *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[8] R. Timofte, V. De Smet, and L. Van Gool, “A+: Adjusted anchored neighborhood regression for fast super-resolution,” in *Proceedings of The 12th Asian Conference on Computer Vision (ACCV’14)*, pp. 111–126, Springer, 2014.

[9] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[10] R. Yaniv, I. John, and M. Peyman, “Raisr: Rapid and accurate image super resolution,” *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 110–125, 2017.

Copyright ©2018 Information Processing Society of Japan.
All Rights Reserved.