深層学習を用いた映像伝送トラフィック削減技術の実験と考察

渡邊 翔太[1]　藤橋 卓也[2]　猿渡 俊介[1]　渡辺 尚[1]

概要：本稿では，ビデオ通話や Web 会議などの映像データのトラヒックを削減する方式として，深層学習による超解像技術を用いた伝送手法を提案する．具体的には，まず送信側においてカメラから得た各ビデオフレームをダウンサンプリングすることで解像度を小さくして送信する．次に受信側では，受け取ったダウンサンプリング後の各ビデオフレームを深層学習を用いて元の画像サイズにアップサンプリングして復元する．また，ダウンサンプリングした各ビデオフレームの送信時には，従来の映像符号化による圧縮を併用することでさらなるデータ量の削減を達成する．実際のビデオ通話時の映像データを用いた性能評価から，提案手法は送信データ量を 90%以上削減しつつ，SSIM (Structual SIMilarity) を指標として用いた画像品質評価で約 0.80 まで画像品質を改善できることが分かった．

1. はじめに

スマートフォンの普及によってコミュニケーションの手段として LINE などの SNS サービスが急速に普及してきた [1]．その中でも，LINE や Skype などによるビデオ通話が頻繁に使われるようになった．ビデオ通話では，電話やメールなどの音声や文字だけでなく，映像情報を利用することができるため，音声あるいは文字だけでは上手く伝えることができなかった情報を補うことができる．例えば，映像情報から相手の表情やしぐさを確認することができるため，コミュニケーションが円滑になる．企業でも，Web 会議やテレビ会議などが導入されはじめている．最近では，カスタマーサポートにビデオ通話を利用することにも注目が集まっている．

一般的に，映像情報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．現在，映像情報を伝送するときには H.264/AVC (Advanced Video Coding) や H.265/HEVC (High Efficiency Video Coding) などの映像符号化技術を用いて映像情報を圧縮する．しかしながら，ビデオ通話の使用率が上昇することに鑑みると，これらの符号化技術だけではなく，ビデオ通話に特化してデータ量のさらなる圧縮を達成する仕組みを実現する必要がある．

このような観点から，本研究では，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超解像技術を用いた手法を提案する．具体的には，まず送信側ではカメラから得た各ビデオフレームをダウンサンプリング（低解像度化）して解像度を落とすことで画像サイズを縮小した後，縮小後のビデオフレーム全体を H.264/AVC などの映像符号化技術で圧縮した上で送信する．次に受信側では，受け取った映像情報から各ビデオフレームを取り出した後，深層学習を用いて各ビデオフレームを元の解像度にアップサンプリング（高解像度化）する．このとき，ビデオ通話をするユーザに特化した生成モデルを深層学習で用いることで，高解像度化後の画像品質を向上させる．実際のビデオ通話時の映像情報を用いて評価した結果，送信データサイズを 90%以上削減しつつ，SSIM (Structual SIMilarity) を指標とした評価で約 0.80 の画像品質を達成した．

本稿の構成は以下の通りである．2 章では，本研究の背景となる既存の技術および関連研究について述べる．3 章では，提案手法を説明する．4 章では，実際のビデオ通話映像を用いた評価を通して提案手法の効果を考察する．最後に 5 章でまとめとする．

2. 関連研究

本研究は映像符号化技術，GAN (Generative Adversarial Network) を用いない超解像技術，GAN を用いた画像処理技術に関係する．超解像技術とは，本来ディスプレイに表示される画像よりもさらに詳細に画像を写しだすために，元の画像から細部を予測する技術である．本来の解像度よりもさらに解像度をあげる技術という点で，本研究は超解像技術の一種と考えることができる．

2.1 映像符号化技術

映像符号化は現在でもビデオ通話に一般的に使用されて

[1] 大阪大学大学院情報科学研究科
[2] 愛媛大学大学院理工学研究科

© 2018 Information Processing Society of Japan

いる技術である．1990 年に策定された国際標準規格 H.261 から現在に至るまで，ITU-T や ISO/IEC が中心となって映像符号化技術は進化を続けている．

2018 年現在における最先端の映像符号化として H.265/HEVC が挙げられる．H.265/HEVC は現在主流の H.264/AVC と比べて約 2 倍の圧縮性能を実現している．H.265/HEVC の利用を通して，4K や 8K と言った超高精細映像を IP 網で伝送すると言った応用が期待されている．

H.265/HEVC はこれまでの映像符号化と同様に，フレーム間予測，量子化，エントロピー符号化を用いて映像情報を圧縮する．1990 年より圧縮効率が進化し続けていることに鑑みると，今後も圧縮効率は向上し続けることが予想される．本稿では，今後も進化が予想される映像符号化と同時に使用可能な超解像技術を，提案システムにおいて併用することで，ビデオ通話に生じるデータ量の大幅な削減を目指す．

2.2 GAN を用いない超解像技術

超解像技術としては，単一画像を対象としたものと，複数画像を対象としたものに分けられる．単一画像を対象とした超解像技術としては，A+[2]，SRCNN[3]，RAISR[4] が研究されている．例えば，A+[2] は補間法を用いることで超解像技術の高速化を実現している．複数画像に対する超解像技術としては，文献 [5], [6] が挙げられる．本研究では，単一画像を対象とした超解像技術に焦点を当てる．複数画像を対象とした超解像技術については議論しない．

深層学習の画像生成モデルとしては，主なものに変分推論モデル，敵対的生成モデル，自己回帰モデルなどがある．変分推論モデルには VAE (Variational Auto Encoder)[7], [8] がある．VAE は，入力された訓練データを正規分布の潜在変数に変換して，その潜在変数を入力として元画像と類似した画像を復元できるモデルである．

自己回帰モデルには，PixelCNN や PixelRNN[9]，Pixel Recursive Super Resolution[10] がある．自己回帰モデルでは，各ピクセルごとに元画像への復元を試みる．Pixel Recursive Super Resolution[10] では，画像認識で利用されるネオコグリトロンモデル [11] に基づく CNN (Convolutional Nural Network)[12] を用いて被写体を大まかに特定した後，PixelCNN を用いることによってピクセルごとに細部まで顔画像を復元できる．より高品質の高解像度画像を生成することができる反面，計算に多大な時間がかかるという欠点がある．

2.3 GAN を用いた画像処理技術

敵対的生成ネットワーク (GAN: Generative Adversarial Network)[13] を画像処理に利用する研究が数多くなされている．GAN は生成モデルと判別モデルの 2 種類のニューラルネットワークを用いる．生成モデルと判別モデルが交互に精度を高め合うことによって，最終的に生成モデルは入力データと非常によく似たデータを高速に生成することができる．画像と画像の相互変換を行っている研究 [14]，1 枚の画像から未来を予測して動画を生成する研究 [15], [16]，2 次元画像から 3 次元画像を生成する研究 [17] など，多様な取り組みがなされている．

GAN を超解像技術に適用する研究もなされている．中でも，DCGAN (Deep Convolutional Generative Adversarial Network) [18] が品質が高い画像を生成できている．DCGAN を超解像に利用した研究としては，LPGAN (Laplacian Pyramid of Generative Adversarial Network)[19], SRGAN (Super-Resolution Using a Generative Adversarial Network)[20] が挙げられる．LPGAN は，GAN がいきなり高解像度の画像を生成することが難しいことを考慮して，GAN を何段にも積み重ねて少しずつ高解像度にしていくことでより高品質な高解像度画像を生成する．SRGAN は，DCGAN の画像認識に利用される VGGNet[21] を損失関数において利用することで超解像技術を行なっている．

本研究と最も関連が深い研究として，人の顔画像に特化して，画像サイズがより小さい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研究である [22]．srez[22] では，DCGAN を利用して顔画像を復元することに成功している．しかしながら，DCGAN を用いた高解像度化技術では人間から見た違和感がない画像が生成できるものの，元の画像と類似してはいるが同一人物とは言い難い顔が生成されることが問題となっている．

3. 提案手法

ビデオ通話における映像情報のデータ量を削減するために，本稿では，送信側でダウンサンプリングした映像情報を圧縮・送信した後，受信側で顔の輪郭情報を損失関数に用いた DCGAN を利用して元映像を高品質に復元する映

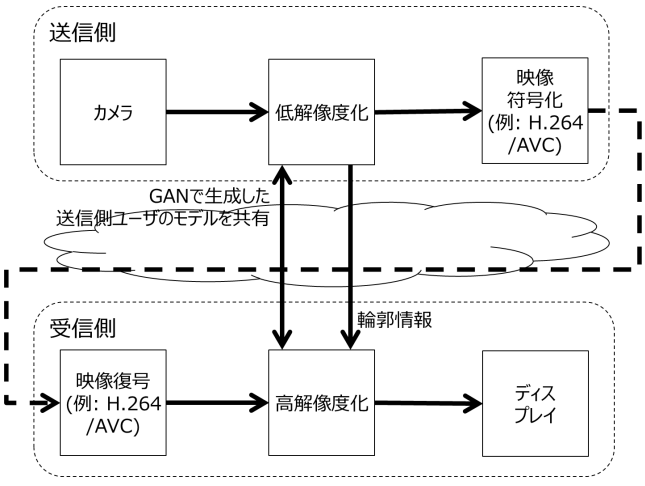

図 1 提案システム概要

© 2018 Information Processing Society of Japan

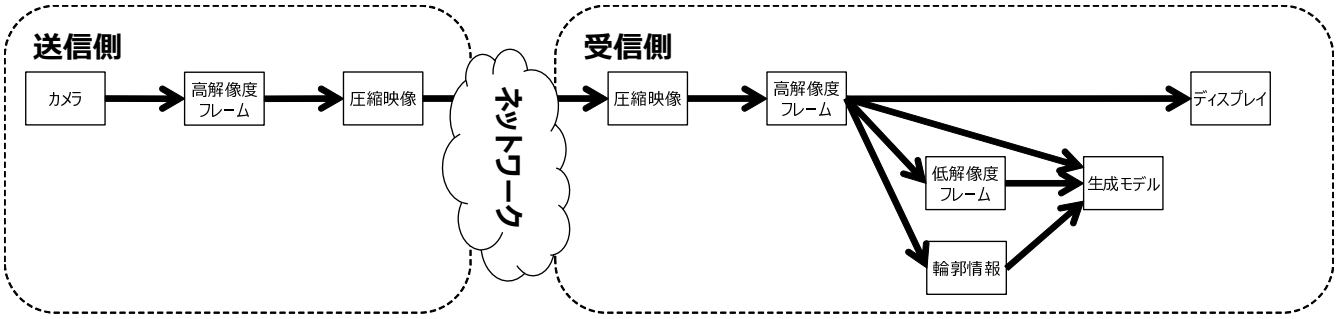

図 2 学習通信フェーズ

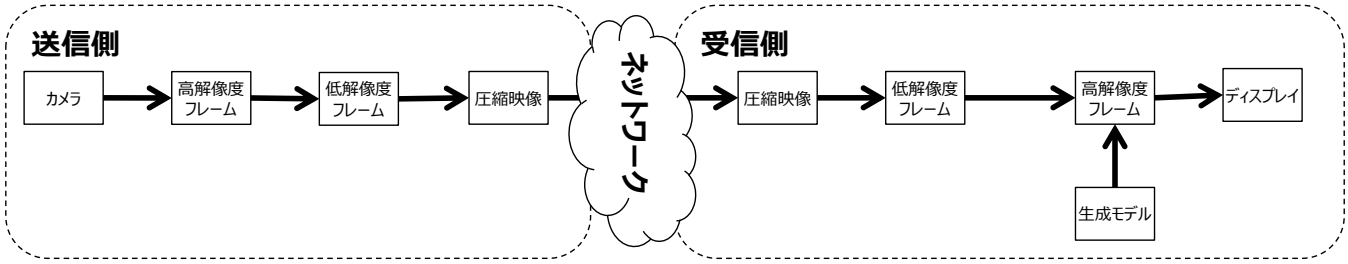

図 3 高解像度化通信フェーズ

像伝送手法を提案する．図 3 に提案システムの基本的なアイディアを示す．提案システムでは，まず，送信側においてカメラから取得した各ビデオフレームに含まれる輪郭情報を抽出すると共に，各ビデオフレームを低解像度化する．その後，H.264/AVC などの映像符号化技術を用いて複数枚の低解像度化したビデオフレームを圧縮して受信側に送信する．受信側では，受け取った符号化映像を復号した後，各ビデオフレームを顔の輪郭情報に基づく生成モデルを用いて高解像度化してからディスプレイに表示する．本システムの特徴として，送信側の低解像度化と受信側の高解像度化のプロセスにおいて，各ビデオフレームにおける顔の輪郭情報を利用する点，DCGAN を用いた生成モデルを送受信端末間で共有する点が挙げられる．

3.1 システム構成

DCGAN を用いた生成モデルの共有方法としては，逐次生成方式と事前共有方式の 2 種類を想定している．

逐次生成方式

逐次生成方式は，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側の学習が終了すると，受信側は学習が終了したことを知らせる信号を送信側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，受信側は高解像度映像のディスプレイ表示と生成モデルの学習を同時に行う．送信側では，カメラから取得した複数枚の高解像度フレームを H.264/AVC などの映像符号化を用いて圧縮してから受信側に送信する．圧縮後の映像情報を受け取った受信側は，まず，受信映像情報を復号することで高解像度フレームを取得してそのままディスプレイに表示する．それと同時に，高解像度フレームから低解像度フレームと顔の輪郭情報を抽出する．最後に低解像度フレーム，顔の輪郭情報を用いて元の高解像度フレームを生成するための生成モデルを学習する．

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，送信側はカメラから取得した各高解像度フレームから低解像度フレームと顔の輪郭情報を生成する．その後，複数枚の低解像度フレームを H.264/AVC などの映像符号化で圧縮してから受信側に送信する．圧縮後の映像情報を受け取った受信側では，受信映像情報を復号することで低解像度フレームを生成する．その後，学習通信フェーズで得られた生成モデル，低解像度フレームから高解像度フレームを生成してディスプレイに表示する．

事前共有方式

事前共有方式とは，送信側ユーザに関する生成モデルをビデオ通話の開始前に受信側に送っておくモデルである．生成モデルを受信側に送信した後は逐次生成方式の高解像度化通信フェーズ (図 3) と同じ動作を行う．事前共有方式を採用することで，逐次生成方式で問題になると予想される学習通信フェーズのオーバヘッドは無くなるものの，生成モデルの学習に用いた画像と異なる状況 (背景や衣服など) でビデオ通話を行った場合に高解像度化された画像の品質が下がる可能性がある．

3.2 DCGAN で用いる生成モデルと判別モデル

本手法では，サイズが $W \times H \times C$ のテンソルである元画像 I_R とその元画像のサイズを $rW \times rH \times C$ のテンソルにダウンサンプリングした画像 I_M を用いる．ここで，

© 2018 Information Processing Society of Japan

W は元画像の幅，H は元画像の高さ，C は画像のチャネル数を表す．また，r はダウンスケーリング係数である．生成モデル $G(= G_{\theta_G}(I_M))$ はダウンサンプリング画像 I_M およびパラメータ θ_G を入力として元画像 I_R と類似する画像を生成するモデルである．このとき，生成モデル G の精度を高めるために生成モデル G のパラメータ θ_G を最適化する．この最適化は N 枚の元画像 $I_R^{(i)}, i = 1, ..., N$ とダウンサンプリング画像 $I_M^{(i)}, i = 1, ..., N$ を用いて以下の式で表される．

$$\hat{\theta}_G = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l(G_{\theta_G}(I_M), I_R)$$

ここで，l は生成モデルの損失関数を表す．生成モデルの損失関数に関しては 3.3 節で詳細に述べる．

生成モデル G のパラメータ最適化のために判別モデル D を用いる．判別モデルは与えられた画像が元画像 I_R による正解データ群と生成モデルが生成した画像 $G_{\theta_G}(I_M)$ による偽データ群のどちらに属するかを分別する 2 値分類器である．生成モデルと判別モデルが互いに精度を高め合うことで学習が行われる．学習はゲーム理論を応用した次のミニマックス法によって最適化する．

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I_R \sim P_{\mathrm{data}}(I)}[\log (D_{\theta_D}(I_R))]$$
$$+\mathbb{E}_{I_M \sim P_{mosaic}(I)}[\log (1 - D_{\theta_D}(G_{\theta_G}(I_M)))]$$

$D_{\theta_D}(I_R)$ は入力画像 I_R が正解データ群に属する確率，$1 - D_{\theta_D}(G_{\theta_G}(I_M))$ は低解像度画像 I_M から生成モデル G によって生成されたデータが偽データ群に属する確率，$\mathbb{E}[\cdot]$ は期待値を意味する．判別モデル D は入力画像と生成モデル G によって生成されたデータを正しく判別するため，これらの確率を最大化するようにパラメータ θ_D を調節する．生成モデル G は自身が生成したデータを判別器 D に正解データとして判断して欲しいため，後者の確率を最小化するようにパラメータ θ_G を調節する．

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，図 3 の高解像度通信フェーズで利用される．図 4, 図 5 に生成器のモデルと判別器のモデルをそれぞれ示す．今回は srez のモデルを利用した．

図 4 の生成モデルでは，1 層目のバッチ正規化層と 2 層目の ReLU (Rectified Linear Unit) 層，3 層目の畳み込み層，4 層目の結合層を 8 層目まで繰り返し，9 層目に 2 倍の Up Scale 層，10 層目にバッチ正規化層，11 層目に ReLU 層，12 層目に転置畳み込み層がある．この 1 層目から 12 層目までをもう一度繰り返して 13 層目から 24 層目とする．さらに，25 層目の畳み込み層と 26 層目に ReLU 層をもう一度繰り返して最後の 29 層目にシグモイド層がある．

図 5 に判別モデルを示す．図 5 に判別モデルでは，1 層目の畳み込み層と 2 層目のバッチ正規化層と 3 層目の ReLU 層の 3 つの層を 18 層目まで繰り返し，19 層目に畳み込み層，20 層目に平均を取る層がある．

バッチ正規化 [23] は，入力データの各要素のバランスが取れていない場合に，値の大きい要素が大きく影響してしまうという問題を解消するための調整する仕組みである．要素ごとに平均 0，分散 1 に正規化している．ReLU[24] は CNN によく利用される活性化関数である [25]. ReLU 関数は以下の式で表される．

$$\mathrm{ReLU}(x) = \max(0, x)$$

シグモイド関数は生物の神経細胞がもつ性質をモデル化した活性化関数である．古くから機械学習に最もよく使用されている関数である．シグモイド関数は以下の式で表される．

$$\mathrm{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

畳み込み層は，CNN に利用される層である．画像の畳み込みは，フィルタの濃淡パターンと類似した濃淡パターンが画像のどこにあるかを抽出する働きがある．深層学習における畳み込み層では，フィルタが重みで構成されているため，抽出したい濃淡構造を学習することになる．

3.3 生成モデルの損失関数

3.2 節に示した生成モデルの損失関数 l は次の式で表される．

$$l = \alpha_1 l_{\mathrm{adversarial}} + \alpha_2 l_{\mathrm{pixel}} + \alpha_3 l_{\mathrm{face}} \qquad (1)$$

ここで α_1, α_2, α_3 は 0 から 1 の実数値をとるパラメータである．$l_{\mathrm{adversarial}}$ は敵対的ニューラルネットワークにおける生成器の損失である．$l_{\mathrm{adversarial}}$ は以下の式で表される．

$$l_{\mathrm{adversarial}} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I_M))$$

l_{pixel} はピクセル単位での損失である，l_{pixel} は元画像をダウンサンプリングした画像 $I_{M_{(x,y)}}$ と生成器が生成した画像をダウンサンプリングした画像 $G_{\theta_G}(I_M)_{(\frac{x}{r}, \frac{y}{r})}$ との画素値の差分の平均をとったものとして以下の式で表される．

$$l_{\mathrm{pixel}} = \frac{1}{r^2 WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} \left| I_{M_{(x,y)}} - G_{\theta_G}(I_M)_{(\frac{x}{r}, \frac{y}{r})} \right|$$

l_{face} は顔の特徴点の損失である．顔の特徴点の検出には HOG (histgram of oriented gradient) 検出器 [26] を利用した [27]．今回は，dlib と iBUG 300-W が公開している既存のデータセット `http://dlib.net/files/shape_predictor_68_face_landmarks.dat.bz2` を用いた．dlib では顔におけるあご，右眉，左眉，鼻，右目，左目，口の合計 68 個の輪郭情報が検出できる．dlib で 68 個の輪郭情報の点の座標を検出する関数を Ψ とすると，l_{face} は次の式で表される．

$$l_{\mathrm{face}} = \frac{1}{68} \sum_{k=1}^{68} ||\Psi(I_R) - \Psi(G_{\theta_G}))||_2$$

© 2018 Information Processing Society of Japan

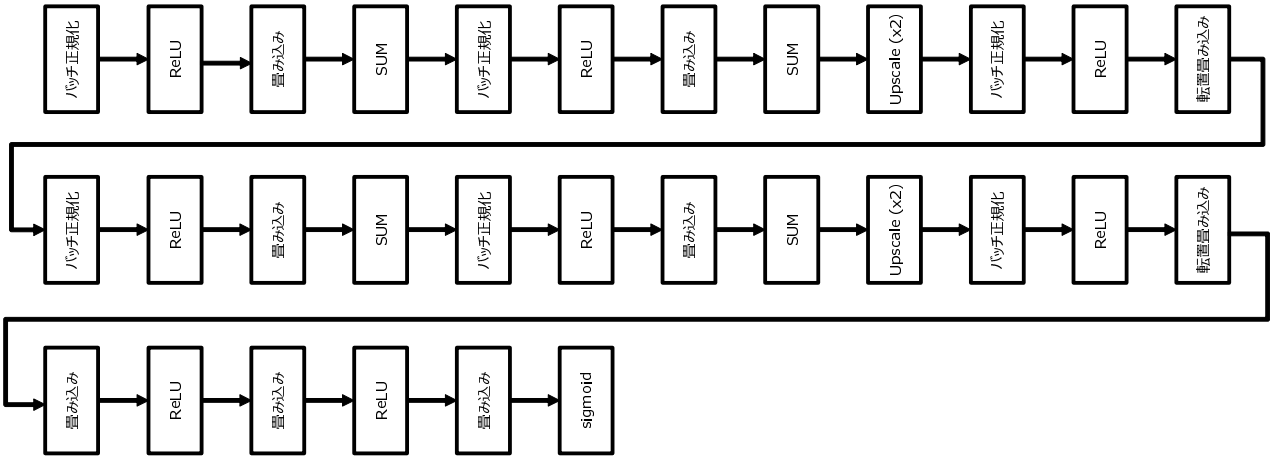

図 4 生成モデル

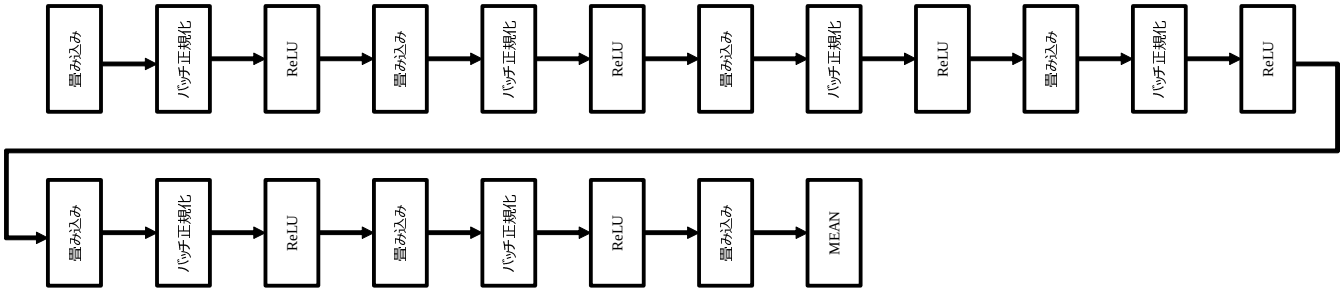

図 5 判別モデル

l_{face} は dlib によって顔の特徴点検出によって顔が検出できた時だけ計算できるため，顔がうまく生成されない学習初期では l_{face} が計算できない．そこで，l_{face} が計算できない顔が検出されない場合は次の式を用いる．

$$l = (1-\alpha)l_{\text{adversarial}} + \alpha l_{\text{pixel}} \qquad (2)$$

ここで，α は 0 から 1 の実数値をとるパラメータである．

4. 性能評価

提案システムによる効果を確認するために，実際のビデオ通話映像を用いて性能評価を行った．

4.1 評価環境

本評価に用いたテスト画像群は，PENTAX KS-2 を用いて撮影した 10 分間の動画から取得した．フレームレートは 30 fps である．撮影した動画の各ビデオフレームは JPEG を用いて圧縮した後，解像度 80×80 画素と 160×160 画素にリサイズすることで 2 種類の入力画像群を作成した．200,000 枚作成した入力画像群から無作為に 16 枚を選択して学習に，同じく無作為に 8 枚を選択して評価に用いた．

OS は Ubuntu 16.04.1，CPU は Intel Xeon プロセッサー E5-2637 v3 を使用した．GPU は使用していない．損失関数のパラメータは $\alpha = 0.90$，学習係数の初期値は 0.0002，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，学習パラメータの更新には Adam [28] を用いた．Adam の初期値は論文 [28] の奨励値を用いている．

4.2 定性評価

本章では，学習が進むにつれて生成モデルによる生成画像がどのように変化するか議論する．図 6 に，解像度 80×80 画素の顔画像を対象としたとき，解像度 160×160 画素の顔画像を対象としたときの生成画像の変化を示す．より具体的には，図 6 では，上段に損失関数として式 (1) と式 (2) を用いたときの解像度 80×80 画素の生成画像，中段に式 (1) のみを用いた解像度 80×80 画素の生成画像，下段に式 (1) のみを用いたときの解像度 160×160 画素の生成画像を示している．このとき，それぞれの評価に用いた 8 枚の画像のうち，2 枚に対する結果を縦に並べている．

また，図 6 では，左の画像から右の画像に向けて学習が進むごとに画像がどのように変化するのかを示している．具体的には，左からモザイク画像，bicubic 補間画像，学習回数が 100 回の生成画像，200 回の生成画像，300 回の生成画像，400 回の生成画像，500 回の生成画像，800 回の生成画像，1,000 回の生成画像，2,000 回の生成画像，3,000 回の生成画像，4,000 回の生成画像，5,000 回の生成画像，8,000 回の生成画像，10,000 回の生成画像，元画像を示している．ここで，モザイク画像とは，対象画像の解像度を縦横それぞれ 1/4 倍にした低解像度画像に Nearest Neighbor 補間を施したものを指す．

© 2018 Information Processing Society of Japan

図 6　学習回数に対する生成画像の変化

図 6 から，生成画像は学習回数が増えるにしたがって元画像に近くなっていることが分かる．学習回数が少ないと人の顔には見えないが，学習回数が 1,000 回を超えてくると元画像により近い品質が達成できていることが分かる．また，損失関数として式 (1) と (2) を用いた上段の生成画像と式 (1) のみを用いた中段の生成画像を比較すると，損失関数として式 (1) と (2) を用いた生成画像の方が式 (1) のみを用いた生成画像よりも，顔の違和感が早くなくなっているため，式 (1) と (2) を用いた方が顔画像生成に適していることがわかる．また，解像度が 80 × 80 画素である中段の生成画像と解像度が 160 × 160 画素である下段の生成画像を比較すると，元の解像度が大きい画像の方がより早く違和感のない画像が生成できることが分かる．これは，解像度が大きい画像の方が，ダウンサンプリング後に画像に残される情報量が多いからであると考えられる．

4.3　データサイズ削減効果

提案システムによって，ビデオ通話時に生じるデータ量をどの程度削減できたか確かめるために，データサイズを評価した．まず，生成モデルのデータサイズは画像サイズに依らず約 92MB であった．事前共有方式で生成モデルを送ったとしても送ることができるサイズであると言える．

図 7 に，評価に使用した画像とそのデータサイズを示す．ここで，図 7 の上段には解像度 80 × 80 画素，下段には解像度 160 × 160 画素を対象としたときの画像を示している．左側から元画像，モザイク画像，生成画像を示すとともに，それぞれの画像の解像度とデータサイズを示している．なお，学習回数は 20,000 回としている．

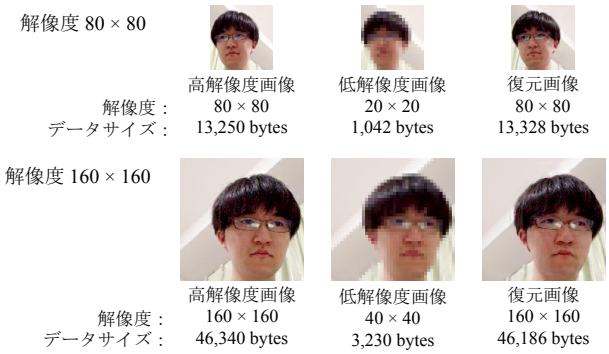

図 7　評価に利用した画像とそのデータサイズ

図 7 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 画素の場合はそれぞれ 13,250 Bytes，1,042 Bytes である．一方で，解像度 160×160 画素の場合は，元画像と低解像度画像のデータサイズがそれぞれ 46,340Bytes，3,230 Bytes であった．どちらの解像度においても，画像情報の伝送に要するデータ量を 90 %以上削減できていることが分かった．

4.4　画像品質

提案システムを用いて得られた生成画像の品質について議論するため，続いて低解像度画像，生成画像および bicubic 補間画像の WPSNR (Weighted Peak Signal-to-Noise Ratio) を比較した．WPSNR とは，YCbCr 色空間 (Y:輝度，Cb:青色系統の色相と彩度，Cr:赤色系統の色相と彩度) 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 [29]．WPSNR を用いたのは，YCbCr 色空間では Y 成分が Cb 成分，Cr 成分よりも多くの情報を所持しているためであ

© 2018 Information Processing Society of Japan

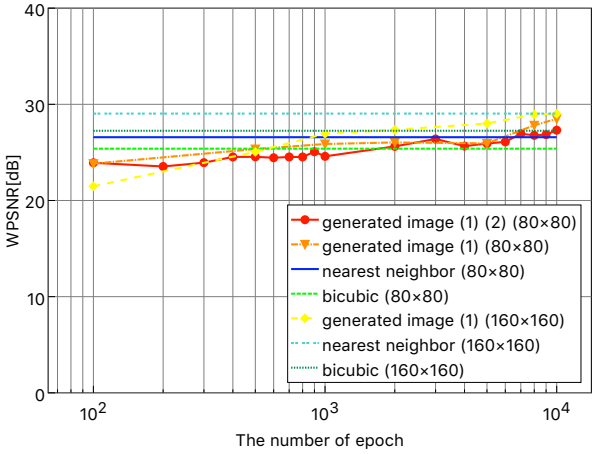

図 8　学習回数と WPSNR との関係

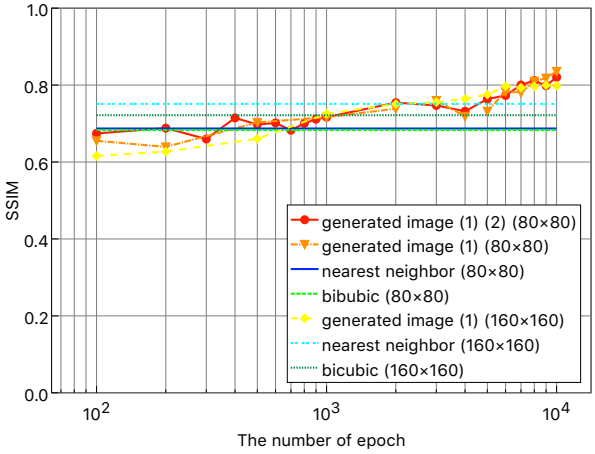

図 9　学習回数と SSIM との関係

る．具体的には，WPSNR を求めるために，以下の式を用いた．

$$\mathrm{PSNR} = -10\log_{10}\frac{\mathrm{MSE}}{255^2}$$

$$\mathrm{MSE} = \frac{1}{nm}\sum_{i=0}^{n}\sum_{j=0}^{m}\{\mathrm{original}(i,j) - \mathrm{encoded}(i,j)\}^2$$

ここで，(n,m) は画像の縦横のピクセル幅，$\mathrm{original}(i,j)$ は高解像度画像におけるピクセル (i,j) の階調値，$\mathrm{encoded}(i,j)$ は復元後画像におけるピクセル (i,j) の階調値を表す．MSE (Mean Squared Error) は元画像と復元後画像の平均二乗誤差を表す．WPSNR は復元画像 8 枚それぞれについて PSNR を計算して，真値で重み付けをしてから画像 8 枚の平均を計算してデシベル値とした．

図 8 に示した評価結果から，次の 2 つのことがわかる．1 つ目は，学習回数が増えるごとに解像度 80×80 画素の元画像に対する生成画像の WPSNR と解像度 160×160 画素の元画像に対する生成画像の 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，解像度が 80×80 画素の画像について，損失関数に式 (1) と (2) の両方を用いた生成画像と式 (1) のみを用いた生成画像の WPSNR の値があまり変わらないことである．WPSNR はピクセルごとの画素値を比較しているため，画像を生成するモデルを評価するのに適していないと考えられる．

次に，SSIM [30] 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 [31]．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 [30] と同じ値 $(K_1 = 0.01, K_2 = 0.03)$ を用いている．さらに文献 [30] と同じく評価前の画像にガウシ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 9 に示した評価結果から，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．また，損失関数に式 (1) と (2) の両方を用いた生成画像と式 (1) のみを用いた生成画像との間で SSIM の値があまり変わらないことが分かる．本稿における性能評価では学習画像数とテスト画像数がともに少ない枚数で行った．生成画像の品質には学習に用いた無作為の 16 枚の画像とテストに用いた 8 枚の無作為の画像に依存するところがあるため，学習と評価に用いる画像の関連性についても考える必要がある．

続いて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 10 に学習通信フェーズにおける学習回数と計算量の関係を示す．図 10 では，解像度 80×80 画素の生成画像と解像度 160×160 画素の生成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 10 から，学習回数が 10,000 回程度を達成するのに，解像度 80×80 画素の画像では 50,000 秒 (約 13 時間)，解像度 160×160 画素の画像では 175,000 秒 (約 48 時間) かかることが分かった．画像の解像度を大きくすると計算に時間がかかるため，解像度の小さい画像を利用する必要があると考えられる．なお，高解像度通信フェーズでは，画像 16 枚を生成するために要した計算時間が約 1.14 秒であった．

4.5　グレースケール画像化によるデータサイズ削減効果

提案システムが必要とする映像情報のデータ量をさらに

© 2018 Information Processing Society of Japan

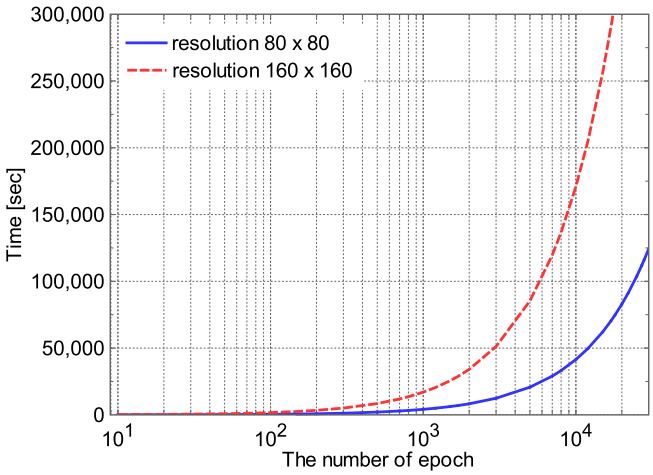

図 10　学習回数と計算量の関係

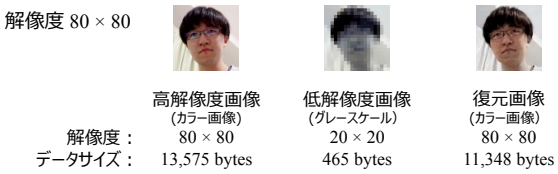

図 11　評価に用いた画像とそのデータサイズ

削減するために，画像をグレースケール化して提案システムを用いた場合の評価も行った．本評価に用いたテスト画像列は，4.1 節のものと同じ動画から取得した．撮影した動画から解像度 80 × 80 の入力画像群を作成した．入力画像群は 10,000 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，無作為に 8 枚を選択してテストに用いた．

本評価は，グレースケールの低解像度画像を高解像度化するために 3.2 節のモデルを用いるとともに，カラー化には [32] を用いた．なお，損失関数は式 (2) を使用する．他の評価パラメータや評価環境は 4.1 節と同じものを用いた．

グレースケール化によってデータ量がどれほど削減できたかを確かめるために，データサイズを評価した．図 11 に評価に使用した画像とそのデータサイズを示す．また，左側からカラーの元画像，グレースケール化とモザイク処理をした低解像度画像，カラー化した生成画像を示している．なお，学習回数は 10,000 回としている．

図 11 における元画像と低解像度画像のデータサイズは，それぞれ元画像が 13,575 Bytes，低解像度画像が 465 Bytes であった．すなわち，画像情報の伝送に要するデータ量を 96 %以上削減できていることが分かる．

5. おわりに

本稿では，ビデオ通話における映像トラヒックを削減する方式として，深層学習による超解像技術を用いた方式を提案した．本提案手法によって，画像品質をある程度保ったままデータ量を大きく削減することが確認できた．

参考文献

[1] LINE 株式会社：2017 年 12 月期通期決算説明会 (Access: 2018/2/10)，https://scdn.line-apps.com/stf/linecorp/ja/ir/all/Q4_earningreleases_JP.pdf (2018).

[2] Timofte, R., De Smet, V. and Van Gool, L.: A+: Adjusted Anchored Neighborhood Regression for Fast Super-Resolution, *In Proceedings of The 12th Asian Conference on Computer Vision (ACCV'14)*, Springer, pp. 111–126 (2014).

[3] Dong, C., Loy, C. C., He, K. and Tang, X.: Image Super-Resolution Using Deep Convolutional Networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 2, pp. 295–307 (2016).

[4] Yaniv, R., John, I. and Peyman, M.: RAISR: Rapid and Accurate Image Super Resolution, *IEEE Transactions on Computational Imaging*, Vol. 3, No. 1, pp. 110–125 (2017).

[5] Borman, S. and Stevenson, R. L.: Super-Resolution from Image Sequences-A Review, *In Proceedings of The IEEE Midwest Symposium on Circuits and Systems 1998*, IEEE, pp. 374–378 (1998).

[6] Farsiu, S., Robinson, M. D., Elad, M. and Milanfar, P.: Fast and Robust Multiframe Super Resolution, *IEEE transactions on image processing*, Vol. 13, No. 10, pp. 1327–1344 (2004).

[7] Kingma, D. P. and Welling, M.: Auto-Encoding Variational Bayes, *In Proceedings of The International Conference on Learning Representations 2014 (ICLR'14)* (2014).

[8] Doersch, C.: Tutorial on Variational Autoencoders, *ArXiv Preprint ArXiv:1606.05908* (2016).

[9] Van Den Oord, A., Kalchbrenner, N. and Kavukcuoglu, K.: Pixel Recurrent Neural Networks, *In Proceedings of The 33rd International Conference on International Conference on Machine Learning 2016 (ICML'16)*, Vol. 48, JMLR.org, pp. 1747–1756 (2016).

[10] Dahl, R., Norouzi, M. and Shlens, J.: Pixel Recursive Super Resolution, *In Proceedings of The IEEE International Conference on Computer Vision (ICCV'17)* (2017).

[11] Fukushima, K.: Neocognitron: A Self-Organizing Neural Network Model for A Mechanism of Pattern Recognition Unaffected by Shift in Position, *Biological Cybernetics*, Vol. 36, pp. 193–202 (1980).

[12] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems 25 (NIPS'12)* (Pereira, F., Burges, C. J. C., Bottou, L. and Weinberger, K. Q., eds.), Curran Associates, Inc., pp. 1097–1105 (2012).

[13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative Adversarial Nets, *In Proceedings of Advances in Neural Information Processing Systems 27 (NIPS'14)*, Curran Associates, Inc., pp. 2672–2680 (2014).

[14] Santhanam, V., Morariu, V. I. and Davis, L. S.: Generalized Deep Image to Image Regression, *In Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition 2017 (CVPR'17)* (2017).

[15] Vondrick, C. and Torralba, A.: Generating the Future with Adversarial Transformers, *In Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition 2017 (CVPR'17)* (2017).

© 2018 Information Processing Society of Japan

[16] Zeng, K.-H., Shen, W. B., Huang, D.-A., Sun, M. and Niebles, J. C.: Visual Forecasting by Imitating Dynamics in Natural Sequences, *In Proceedings of The IEEE International Conference on Computer Vision 2017 (ICCV'17)*, pp. 3018–3027 (2017).

[17] Kehl, W., Manhardt, F., Tombari, F., Ilic, S. and Navab, N.: SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again, *In Proceedings of The IEEE International Conference on Computer Vision 2017 (ICCV'17)*, pp. 1530–1538 (2017).

[18] Radford, A., Metz, L. and Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, *In Proceedings of International Conference on Learning Representations 2016 (ICLR'16)* (2016).

[19] Denton, E. L., Chintala, S., szlam, a. and Fergus, R.: Deep Generative Image Models Using a Laplacian Pyramid of Adversarial Networks, *Advances in Neural Information Processing Systems 28* (Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. and Garnett, R., eds.), Curran Associates, Inc., pp. 1486–1494 (2015).

[20] Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. and Shi, W.: Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network, *In Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition 2017 (CVPR'17)* (2017).

[21] K, S. and A, Z.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *In Proceedings of The International Conference on Learning Representations 2015 (ICLR'15)* (2015).

[22] Garcia, D.: Image Super-Resolution Through Deep Learning (Access: 2017/12/21), https://github.com/david-gpu/srez (2016).

[23] Ioffe, S. and Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, *In Proceedings of The 17th International conference on machine learning (ICMI'15)*, pp. 448–456 (2015).

[24] Glorot, X., Bordes, A. and Bengio, Y.: Deep Sparse Rectifier Neural Networks, *In Proceedings of The 14th International Conference on Artificial Intelligence and Statistics*, pp. 315–323 (2011).

[25] Jarrett, K., Kavukcuoglu, K., Ranzato, M. and LeCun, Y.: What Is The Best Multi-Stage Architecture for Object Recognition?, *In Proceedings of The IEEE International Conference on Computer Vision (ICCV'09)*, IEEE, pp. 2146–2153 (2009).

[26] Dalal, N. and Triggs, B.: Histograms of oriented gradients for human detection, *In Proceedings of The IEEE Computer Society Conference onComputer Vision and Pattern Recognition 2005 (CVPR'05)*, Vol. 1, IEEE, pp. 886–893 (2005).

[27] King, D. E.: Max-Margin Object Detection (Access: 2018/5/1) (2015).

[28] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *In Proceedings of The International Conference on Learning Representations 2015 (ICLR'15)* (2015).

[29] Wang, Z., Lu, L. and Bovik, A. C.: Video Quality Assessment Based on Structural Distortion Measurement, *Signal Processing: Image Communication*, Vol. 19, No. 2, pp. 121–132 (2004).

[30] Wang, Z., Bovik, A. C., Sheikh, H. R. and Simoncelli, E. P.: Image Quality Assessment: From Error Visibility to Structural Similarity, *IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp. 600–612 (2004).

[31] Wang, Z. and Bovik, A. C.: 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).

[32] Satishi, I., Edgar, S.-S. and Hiroshi, I.: Let there be Color!: Automatic Colorization of Grayscale Images (Access: 2018/4/10), https://github.com/satoshiiizuka/siggraph2016_colorization (2017).

© 2018 Information Processing Society of Japan