修 士 論 文

題 目

動画配信におけるフレームの特徴量に基づく映像の超解像処理手法の提案

指導教員

報 告 者

大石　貴之

岡山大学大学院自然科学研究科電子情報システム工学専攻

令和 3 年 1 月 28 日 提出

要約

近年，多くの利用者が動画配信サービスでコンテンツを視聴している．動画配信サービスに対する利用者の満足度は，サーバの配信速度やクライアントの受信速度といったサーバとクライアントとの間の通信環境に大きく依存している．サーバとクライアント間の通信状況が悪い場合，クライアントは動画の再生中に中断が発生する可能性がある．この再生中断を減らすため，通信状況に応じて配信動画の品質を変更する手法が提案されているが，受信映像が低解像度化して視聴品質が低下する問題がある．

低品質の映像を受信した場合，受信映像の各フレームに対して解像度が向上する超解像処理を行うことで，クライアントは高品質の映像に変換して再生できる．しかし，クライアント計算機において，CPU やメモリといった計算リソースの性能が十分でない場合，受信した映像を構成するすべてのフレームに対して，リアルタイムに超解像処理を行うことは難しい．また，特徴量の多少に関わらずすべてのフレームに対して超解像処理を行うため，視覚的な映像品質の向上率は小さくなり，効率が悪い．

本論文では，低品質な映像の受信時に特徴量が多いフレームを優先して超解像処理を行いながら動画を再生する手法を提案する．提案手法では，クライアントは一定時間分の映像をバッファリングしながら再生する．このとき，バッファリング中の映像は，再生が開始されるまでの間で，特徴量が多く視覚的な品質向上の効果が高いと予測されるフレームを優先して超解像処理を行うことで，視覚的な映像品質を向上させる．フレームの知覚的類似性を用いた映像品質の評価では，提案手法は，特徴量に応じてフレームを選択しない手法，および超解像処理を行わない手法とそれぞれ比較して，視覚的な品質が向上することを示した．

目次

図 目 次

表 目 次

第 1 章

はじめに

近年，動画配信サービスの普及により全世界のビデオトラフィックが急増しており [1]，通信環境の変化に適応した動画配信システムが必要となっている．サーバとクライアントとの間の通信状況が悪い場合，クライアントは動画の受信が再生に追いつかず，再生中に中断が発生する可能性がある．この再生中断を減らすため，Adaptive Bitrate と呼ばれる配信方式 [2, 3] が提案されており，多くの動画配信サービスで採用されている．Adaptive Bitrate では，クライアントは通信状況に応じて受信する映像の品質を切り替えることで，再生中断の発生を抑制できる．一方で，通信状況が悪い場合に受信する映像の品質が低下する問題がある．

この問題を解決する方法として，低品質映像を受信した場合において，受信映像の各フレームに対して超解像処理を行う方式が挙げられる．超解像とは，低解像度画像から高解像度画像を予測して生成する技術である．低解像度の映像を受信した場合において，受信映像のフレームに超解像処理を行い，高解像度のフレームを生成して再生することで，視聴動画の品質が向上する．しかし，高精度な超解像処理は計算量が多いため，クライアント計算機における CPU やメモリといった計算リソースが不足している場合，受信した映像を構成するすべてのフレームに対してリアルタイムに超解像処理を行うことは難しい．

バッファリング映像に対して超解像処理を行う方式では，クライアントがバッファリングして再生開始を待つ状態のフレームに対して超解像処理を行う．これにより，再生が開始されるまでの間に可能な限りのフレームを高解像度に変換できる．また，特徴量が少ないフレームに対して超解像処理を行った場合，視覚的な映像品質の向上率は小さく，効率が悪い．このため，特徴量が多いフレームに対して優先的に超解像処理を行うことで，再生映像の視覚的な品質を向上できる．

本論文では，低品質な映像の受信時に特徴量が多いフレームを優先して超解像処理を行いながら映像を再生する手法を提案する．提案手法では，クライアントは一定時間分の映像をバッファリングしながら再生する．このとき，バッファリング中の映像は，再生が開始されるまでの間で，特徴量が多く視覚的な品質向上の効果が高いと予測されるフレームを優先して超解像処理を行うことで，視覚的な映像品質が向上する．

第 2 章

画像の拡大技術

2.1　画素補間

画像を拡大して表示する場合，原画像を拡大した画像（以下，拡大画像）を生成する必要がある．これは，連続画像である映像の表示においても同様である．拡大画像は原画像に比べて多くの画素値をもつため，原画像では存在しない画素値を補間する必要がある．画素値の補間では，ニアレストネイバ法，バイリニア法，およびバイキュービック法 [4] といった手法が主に利用され，補間画素周辺の画素情報を基に，補間画素の画素値を求める．

図 2.1 に示されている鳥を写した原画像の矩形領域に対して，ニアレストネイバ法，バイリニア法，およびバイキュービック法を用いてそれぞれ 4 倍に拡大した画像を図 2.2 に示す．ニアレストネイバ法では，補間画素に最も近い位置に存在する画素値を補間画素の画素値に設定して補間する．ニアレストネイバ法による補間は，補間処理が容易であるとともに，原画像の画素値を失わない利点がある．しかし，周辺画素の画素値をこのまま補間画素として利用するため，エッジにジャギーが発生する．

バイリニア法では，補間画素の周辺 4 画素を基に，縦横双方向から直線的に補間して画素値を求める．バイリニア法による補間は，周辺画素の平均化であるため，ニアレストネイバ法に比べてエッジは滑らかになる一方で，高周波成分を生成できず，画像にぼやけが発生する．

バイキュービック法では，補間画素の周辺 16 画素を基に，縦横双方向から 3 次式で補間して画素値を求める．バイキュービック法による補間は，バイリニア法と同様に，エッジが滑らかになる．また，バイリニア法に比べてぼやけの発生を抑制できる．しかし，補間が周辺画素の平均化である点はバイリニア法と同様であり，高周波成分を生成できないため，エッジをシャープに保つことはできない．

図　2.1　鳥の原画像

2.2　超解像

画像の拡大時に，高周波成分を推定して高解像度化する超解像技術が研究されている．超解像では，2.1 節で挙げた一般的な画素補間による拡大とは異なり，画像の特性や前提知識を基に高解像度化を行う．

主な超解像手法は，複数枚の類似画像を基に一枚の高解像度画像を生成する再構成型超解像，および学習用画像を用いて高画質画像と低画質画像の対応パターンを学習する学習型超解像の 2 種類に分類される．近年，学習型超解像では，畳み込みニューラルネットワーク（以下，CNN）を用いた手法が従来手法に比べ高精度に超解像を行うことができ，多くの学習モデルが提案されている．

図 2.1 に示す画像の矩形領域に対して，畳み込みニューラルネットワークを用いた超解像モデルである SRCNN [5] を用いて，4 倍に拡大画像を図 2.3 に示す．図 2.2 と図 2.3 を比較すると，SRCNN による拡大では，他の 3 種類の手法と比較して，エッジがシャープになっている．SRCNN は 3 層の CNN モデルであるが，従来の CNN を用いない学習型超解像手法に比べて高精度な超解像が可能である．最近では，SRCNN の他に，高精度な CNN 超解像モデルとして，SRCNN より畳み込み層の多いモデル [6]，および敵対的生成ネットワークを用いたモデル [7] が提案されている．

映像は連続した画像 (フレーム) であるため，各フレームに対して 2.2 節で示した単一画像の超解像手法を適用することで，映像の超解像が可能である．しかし，映像超解像の品質は

図　2.2　3 種類の手法による鳥の拡大画像

図　2.3　SRCNN による鳥の拡大画像

フレームごとの超解像精度のみでなく，各フレーム間の動きの整合性も重要となる．そこで，フレーム間の動きに一貫性がある映像超解像を行う手法 [8,9] が提案されており，高精度な映像超解像が可能である．

2.3　動画配信における超解像の利用

動画配信サービスの利用時に低解像度映像を受信した場合，受信映像に超解像を適用することで高解像度の映像を再生できる．低解像度映像を受信する状況として，低解像度映像のみが配信されている場合，および Adaptive Bitrate 配信において通信状況が悪い場合が考えられる．しかし，配信動画の再生に超解像を行うため，受信する各フレームに対してリアルタイムに超解像を行う必要があり，CPU やメモリといった計算リソースが不足している場

合は難しい．そこで，リアルタイムに映像超解像を行う手法が提案されている．

論文 [10] では，映像の圧縮に着目した手法を提案している．この手法では，Group Of Picture におけるキーフレームにのみ超解像を行うことで，超解像されたフレームを用いて復号される他のフレームに超解像が伝播し，すべてのフレームで超解像を行うことができる．しかし，この手法は映像の符号化に依存しており，MotionJPEG といったフレーム間予測を行わない符号化を用いる場合は利用できない．また，キーフレームのみに対して超解像を行うため，超解像されたキーフレームを基に生成されるフレームの超解像精度は，フレームに超解像を直接適用した場合に比べて低くなる．

論文 [11] では，計算リソースに応じて深度を変更可能な深層 CNN モデルを提案している．この手法を用いることで，計算リソースに応じた深度のモデルでリアルタイムに超解像を行いながら動画を再生できる．しかし，この手法の想定環境は，軽量な CNN モデルにおいてリアルタイムに超解像処理が可能な状況であり，軽量な CNN モデルにおいてリアルタイムに超解像を行うことができない環境を想定していない．

第 3 章

超解像の評価指標

3.1　概要

超解像によって生成された画像の品質を定量的に評価することは難しい．このため，高解像度画像を低解像度化して超解像により復元した画像と元の高解像度画像との類似度を超解像の精度とすることが一般的である．本章では，本論文で超解像映像の評価に用いる 3 種類の代表的な画像類似性の指標について説明する．

3.2　Peak Signal-to-Noise Ratio (PSNR)

PSNR は，画素値の最大値と各画素値の平均誤差との比率を示す．各画素値の最大値を MAX_I，各画素値の差の二乗の平均値を MSE で表すと，PSNR は式 3.1 で表される．

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \tag{3.1}$$

各ピクセルの誤差が小さいほど MSE の値は小さくなり PSNR の値は大きくなるため，PSNR が大きいほど画像の類似度が高いと判断できる．しかし，PSNR は対応する画素値を単純に比較するため，画素に対してずれが発生すると評価値は大きく低下し，人の知覚特性と異なる評価となる．

3.3　Structural Similarity Index Measure (SSIM)

SSIM [12] では，各画素の類似度について，評価する画素および周辺画素の画素値を基に計算する．評価する各画素の座標をそれぞれ $\boldsymbol{x}$，$\boldsymbol{y}$ とし，各座標の周辺 $N * N$ 画素の平均画素値を μ_x，μ_y，標準偏差を σ_x，σ_y，共分散を σ_{xy} とすると，SSIM は式 3.2 で表される．なお，C_1，C_2 は，分母がゼロに近づいた場合に値を安定させる定数である．

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.2}$$

すべての画素において，式 3.2 を用いて計算した SSIM の平均値が類似度の評価指標となる．SSIM は 0 から 1 の値をとり，値が大きいほど類似度が高い評価となる．SSIM では，画素ごとの画素値を単純に比較せず，周辺領域の画素値を利用するため，PSNR に比べて画素のずれに応じた評価値の低下は小さい．しかし，SSIM はスケーリングや回転といった幾何学的歪みの影響を受けやすく [13]，人の知覚特性とは異なる場合がある．

3.4　Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS [14] は，人の知覚的類似性を学習させたニューラルネットワークによる評価値である．様々な歪みを与えた画像の知覚的類似性に対して多くのユーザが類似性を評価したデータをニューラルネットワークに学習させており，PSNR や SSIM に比べ歪み耐性が高く，人の知覚に沿って類似性を評価できる．LPIPS の値が小さいほど，画像の知覚的類似度は高い評価となる．

3.5　各指標による類似度の評価

図 3.1 に，部屋を写した原画像，部屋をズームした画像，および部屋にぼかしを入れた画像の 3 種類をそれぞれ示す．また，表 3.1 に，図 3.1 のズーム画像およびぼかし画像について，原画像との類似度の評価値をそれぞれ示す．

表 3.1 より，PSNR および SSIM における原画像との類似度は，ズーム画像に比べてぼかし画像の方が高い．一方，LPIPS における原画像との類似度は，ぼかし画像に比べズーム画像の方が高い．

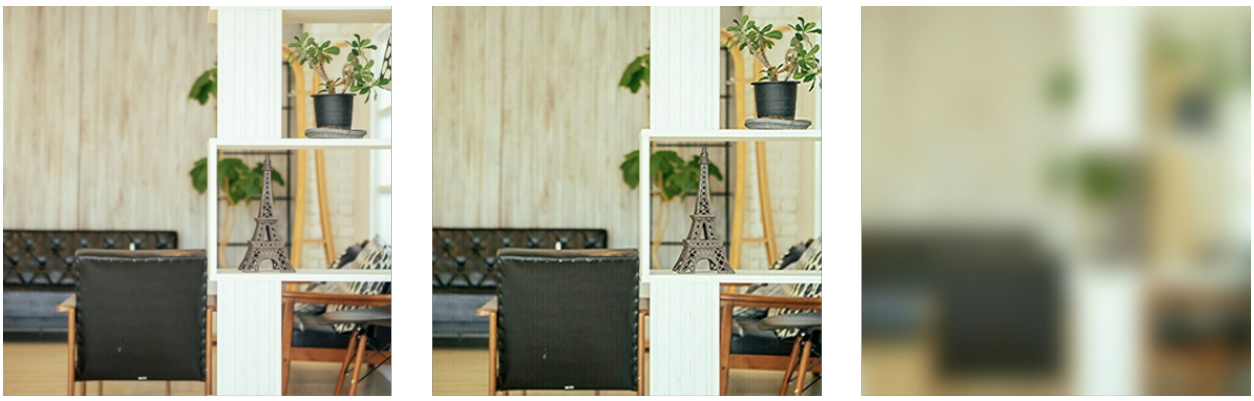

図　3.1　部屋の原画像と変換した画像

表　3.1　部屋画像の各評価値による原画像との類似度

	ズーム画像	ぼかし画像
PSNR	15.175	18.305
SSIM	0.477	0.621
LPIPS	0.299	0.694

第 4 章

特徴量と超解像精度の関連

4.1　特徴量検出

特徴量検出に関する研究では，Features from Accelerated Segment Test（FAST） [15] や Accelerated KAZE（A-KAZE） [16] といった高速なコーナー特徴検出手法が提案され，顔認識や Simultaneous Localization and Mapping（SLAM）といったリアルタイム処理に利用されている [17,18].

街を写した原画像，および街の画像に対して FAST を用いて検出したコーナーを描画した画像を図 4.1 にそれぞれ示す．また，空を写した原画像，および空の画像に対して FAST を用いて検出したコーナーを描画した画像を図 4.2 にそれぞれ示す．街の画像は，建物や車といった多くの物体が存在する複雑な画像であり，検出されるコーナーは 5,407 個である．空の画像は，空と雲のみが写った単純な画像であり，検出されるコーナーは 30 個である．

4.2　特徴量と超解像の関連

図 4.1 および図 4.2 の原画像を 0.25 倍に縮小し，バイキュービック法および SRCNN で 4 倍に復元した画像を図 4.3，図 4.4 にそれぞれ示す．また，図 4.3 の 2 画像と図 4.1 の原画像との類似度，および図 4.4 の 2 画像と図 4.2 の原画像の類似度について，評価結果を表 4.1 に示す.

表 4.1 より，街の画像では，SRCNN を用いた復元画像と原画像との類似度は，バイキュービック法を用いた復元画像と原画像との類似度に比べ，3 種類の評価値すべてで高い．一方で，空の画像では，SRCNN を用いた復元画像と原画像との類似度は，バイキュービック法

図　4.1　街の原画像およびコーナーを描画した画像

表　4.1　街画像および空画像における手法ごとの復元画像と原画像の類似度評価

	街 (バイキュービック)	街 (SRCNN)	空 (バイキュービック)	空 (SRCNN)
PSNR	22.421	22.689	40.186	39.660
SSIM	0.615	0.628	0.954	0.948
LPIPS	0.561	0.487	0.188	0.182

を用いた復元画像と原画像との類似度に比べ，LPIPS の評価において高い一方で，PSNR および SSIM の評価において低く，評価手法に応じて結果が異なる．

バイキュービック法および SRCNN で比べたとき，LPIPS による評価では，街の画像における差は 0.74 となる．一方で，空の画像における差は 0.006 となり，小さい．

次に，街の画像について，図 4.3 における画像の一部領域を拡大した画像を図 4.5 に示す．図 4.5 より，コーナー部では SRCNN を用いた超解像によるシャープ化の効果が大きい．このため，コーナーが多い街の画像では，超解像による効果は大きい．

空の画像について，コーナーが少ない画像では，画素補間によるぼやけの発生が少なく，超解像による視覚的な品質向上の効果が小さいため，LPIPS の差は小さい．

図　4.2　空の原画像およびコーナーを描画した画像

以上より，コーナーの特徴が多い画像について，超解像で拡大した場合における精度向上効果は高い．

バイキュービック法

SRCNN

図　4.3　縮小した街の原画像の各手法による拡大画像

バイキュービック法

SRCNN

図　4.4　縮小した空の原画像の各手法による拡大画像

図　4.5　図 4.3 の一部領域（赤枠）を拡大した画像

第 5 章

提案手法

5.1　概要

本研究では，低品質な映像の受信時に特徴量が多いフレームを優先して超解像処理を行いながら動画を再生する手法を提案する．提案手法は，リアルタイムに超解像を行うことができない環境を想定し，バッファリングしている映像に対して，特徴量が多く超解像による視覚的な品質向上の効果が高いと予測されるフレームを優先して超解像処理を行い，再生する映像のフレームの視覚的な品質を向上する．本章では，提案手法におけるバッファリング映像の超解像処理および超解像フレームの優先順位付けについて述べる．

5.2　バッファリング映像の超解像処理

多くの動画配信システムの利用時に，クライアントは再生中に中断が発生しないようにするため，一定のデータを計算機にバッファリングしながら動画を再生する．このとき，リアルタイムに超解像処理を行うことができない場合，フレームに超解像を適用しながら映像を再生できない．そこで，提案手法では，再生前に受信が完了しているバッファリング映像に対して超解像処理を行うことで，リアルタイムに超解像を行うことができない環境においても同様に，再生が開始されるまでの間で，一部のフレームに超解像処理を行うことができる．

5.3　超解像フレームの選択手法

バッファリング映像のフレームに対して，特徴量に基づき優先して超解像を行うフレームを選択する手法を提案する．提案手法の処理手順は以下の通りである．

1. バッファリングフレームのバッチ分割

2. 各バッチのフレームにおけるコーナーの総数の算出

3. 各バッチのコーナー総数による順位付け

はじめに，バッファリング中のフレームを N 枚ごとのバッチに分割する．Adaptive Bitrate による動画配信では，再生映像の解像度の切り替え頻度が高い場合，視聴品質が低下する [19]．本手法における再生映像でも同様に，視聴するフレームの拡大手法が頻繁に変化すると，解像度の変化により視聴品質が低下することが考えられる．そこで，N 枚ごとのバッチに分割し，各バッチ内のフレームは同一の手法で拡大する．これにより，連続した N 枚以上のフレームが同一の手法で拡大されるため，再生するフレームにおける拡大手法の頻繁な変化による視聴品質の低下を抑制できる．

次に，各バッチのフレームにおけるコーナーの総数を算出する．4.2 節で述べたように，コーナー特徴が多いフレームは，超解像で拡大した場合における精度向上効果が高い．そこで，超解像を行うフレームを選択する指標として各バッチのフレームにおけるコーナーの総数を用いる．なお，コーナー検出のアルゴリズムは，コーナーの高速検出が可能な FAST [15] を用いる．

最後に，各バッチのコーナーの総数による順位付けを行う．コーナーの総数が多いバッチから順番に，超解像を行う．

第 6 章

実装

提案手法において，超解像を行いながら受信映像を再生するプレイヤの構成を図 6.1 に示す．プレイヤは，受信モジュール，超解像モジュール，および再生モジュールの 3 種類で構成される．

受信モジュールは，配信サーバから S 秒分の映像フレーム (以下，セグメント) を受信してバッファに保存し，2 個分のセグメントを保存すると受信を停止する．次に，再生モジュールから受信要求を受けると，新たなセグメントを配信サーバから受信する．再生モジュールは，1 個のセグメントの再生が終了した後に受信を要求する．このため，受信モジュールは，再生済みのセグメントを棄却して，新たに受信したセグメントをバッファに保存することで，バッファに 2 個分のセグメントが常に保存された状態とする．なお，受信モジュールと配信サーバの通信プロトコルは，Adaptive Bitrate 配信で主に利用される HTTP を用いる．

超解像モジュールでは，バッファに保存されており次に再生する予定のセグメントに対して，5 章で述べた手法で超解像を行う．超解像を行っているセグメントの再生が開始されると，このセグメントの超解像を終了し，バッファに保存されており次に再生するセグメントに対して超解像を開始する．なお，超解像では，CNN を用いた学習型超解像モデルである FSRCNN [20] を用いて，映像フレームを 4 倍に拡大する．また，超解像を行うことができないフレームは，バイキュービック法で 4 倍に拡大する．

再生モジュールでは，バッファに保存されている映像を再生する．1 個のセグメントの再生が終了すると，次のセグメントの再生を開始し，受信モジュールに新たなセグメントの受信を要求する．

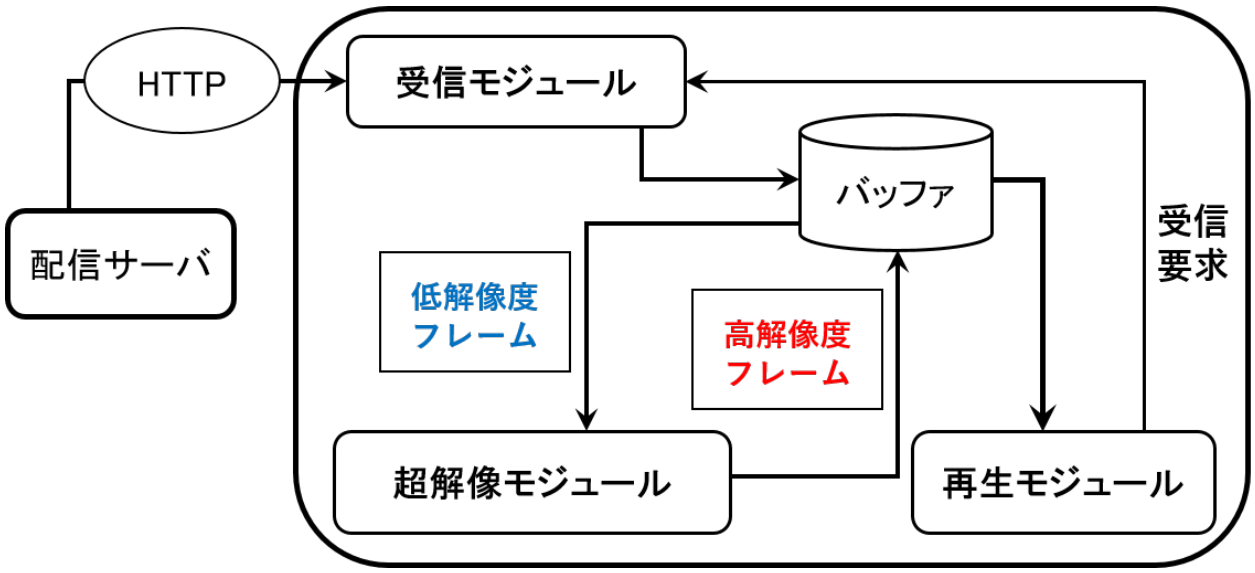

図　6.1　提案手法における映像再生プレイヤの構成

第 7 章

実験評価

7.1　評価環境

6 章で示したプレイヤを実装した計算機と動画配信を行うサーバを Ethernet で接続し，提案手法を評価する．動画配信のサーバソフトウェアは，Apache HTTP Server [21] を用いた．評価に用いた計算機の性能を表 7.1 に示す．クライアントとサーバは，映像の再生に十分な速度で通信可能である．また，クライアントは，映像の再生を開始すると，最後まで再生する．

7.2　評価に用いる映像

評価に用いる 3 種類の映像を，表 7.2 に示す．すべての映像は，開始から 10 分間をトリミングして用いる．

Tears of Steel [22] は，実写と CG が混在し，フレームの時間的変化が大きい映像である．また，他の 2 種類の映像とアスペクト比を揃えるため，左右を切り取りアスペクト比を 16:9 にクロップした映像を用いる．Big Buck Bunny [23] は，アニメーション映像である．Herzmark Homestead [24] は，ドローンで森を空中から映し続けた映像であり，フレームの時間的変化が小さい．

7.3　映像の種類による映像品質への影響

提案手法，前方のフレームを優先して超解像処理を行う単純手法，および全フレームをバイキュービック法で拡大するバイキュービック手法の 3 種類について，再生された全フレー

表 7.1 計算機の性能

Server	CPU	Intel®Pentium(R) CPU G4400 (3.30 GHz) × 2
	Memory	7.7 GBytes
	OS	Ubuntu 18.04.1 LTS
Client	CPU	Intel®CORE(TM) i5-7500 CPU (3.40 GHz)
	Memory	7.9 GBytes
	OS	Windows 10 Pro

表 7.2 評価に用いる映像の構成

タイトル	動画時間	解像度
Tears of Steel [22]	10 分	144 x 256 pixel (144p) 180 x 320 pixel (180p) 270 x 480 pixel (270p)
Big Buck Bunny [23]	10 分	144 x 256 pixel (144p) 180 x 320 pixel (180p) 270 x 480 pixel (270p)
Herzmark Homestead [24]	10 分	144 x 256 pixel (144p) 180 x 320 pixel (180p) 270 x 480 pixel (270p)

ムにおける平均品質を評価する．評価に用いる映像のフレームレートは 24fps であり，各セグメントの映像時間を 20 秒，各バッチのフレーム数は 10 枚とする．評価項目は，平均 PSNR，平均 SSIM，および平均 LPIPS の 3 種類である．

9 種類の映像で評価を行った結果を表 7.3 に示す．表 7.3 より，Tears of Steel および Big Buck Bunny の場合，すべての解像度の映像におけるすべての評価項目について，提案手法が最も高い．一方で，Helzmark Homestead の場合，180p の映像における平均 SSIM 以外の評価項目について，単純手法が最も高い．また，バイキュービック手法は，すべての評価項目について提案手法に比べて低い．一方で，Tears of Steel の 270p の映像，および Big Buck Bunny の 144p と 180p の映像では，バイキュービック手法は単純手法に比べて高い．

表 7.3 各視聴映像の品質評価

受信映像		提案手法			単純手法			バイキュービック手法		
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Tears of Steel	144p	**29.851**	**0.843**	**0.222**	29.799	0.841	0.227	29.695	0.837	0.272
	180p	**31.214**	**0.861**	**0.210**	31.144	0.859	0.217	31.122	0.857	0.240
	270p	**32.730**	**0.885**	**0.187**	32.660	0.884	0.193	32.702	0.884	0.198
Big Buck Bunny	144p	**28.152**	**0.799**	**0.271**	27.452	0.795	0.279	28.101	0.790	0.321
	180p	**29.075**	**0.818**	**0.257**	28.203	0.815	0.268	29.045	0.811	0.290
	270p	**30.176**	**0.850**	**0.223**	30.152	0.848	0.231	30.151	0.848	0.238
Helzmark Homestead	144p	20.478	0.439	0.561	**20.484**	**0.439**	**0.557**	20.377	0.407	0.608
	180p	20.278	**0.433**	0.570	**20.308**	0.432	**0.567**	20.216	0.412	0.596
	270p	20.013	0.434	0.560	**20.039**	**0.435**	**0.559**	19.982	0.425	0.572

7.4　映像の解像度による超解像フレーム数への影響

提案手法と単純手法において，超解像処理が行われたフレーム数を評価する．評価に用いる映像のフレームレートは 24fps であり，各セグメントの映像時間を 20 秒，各バッチのフレーム数は 10 枚とする．

3 種類の映像で評価した結果を表 7.4 に示す．表 7.4 より，提案手法および単純手法において，受信する映像の解像度が高いほど超解像が行われたフレーム数は少ない．また，提案手法で超解像が行われたフレーム数は，単純手法と比べて，144p の映像では平均して約 609 フレーム，180p の映像では平均して約 225 フレーム，270p の映像では平均して約 100 フレーム少ない．

7.5　映像のフレームレートによる映像品質への影響

提案手法，単純手法，およびバイキュービック手法において，フレームレートが異なる 3 種類の映像を再生した場合の視聴品質を評価する．評価には，解像度が 144p，フレームレートが 24fps，30fps，60fps の 3 種類の Big Buck Bunny を用いる．また，各セグメントの映像時間を 20 秒，各バッチのフレーム数は 10 枚とする．評価項目は，平均 PSNR，平均 SSIM，平均 LPIPS の 3 種類である．

3 種類のフレームレートの映像で評価を行った結果を表 7.5 に示す．表 7.5 より，すべてのフレームレートの映像におけるすべての評価項目において，提案手法が最も高い．また，

表 7.4 視聴映像の超解像フレーム数

受信映像		提案手法の超解像フレーム数	単純手法の超解像フレーム数
Tears of Steel	144p	10,388	**10,778**
	180p	6,409	**6,541**
	270p	2,659	**2,757**
Big Buck Bunny	144p	10,101	**10,907**
	180p	6,360	**6,658**
	270p	2,597	**2,728**
Herzmark Homestead	144p	9,568	**10,150**
	180p	5,868	**6,112**
	270p	2,382	**2,452**

SSIM および LPIPS による評価において，提案手法と単純手法は，フレームレートが高くなると評価は低くなる．しかし，PSNR による評価では，提案手法ではフレームレートが高くなると評価が低くなる一方で，単純手法では，24fps および 30fps の映像に比べて 60fps の映像における評価が高い．

7.6　バッチのフレーム数による映像の視聴品質への影響

提案手法において，バッチのフレーム数の変化に応じて，再生映像における拡大手法の変化回数および映像品質を評価する．評価には，解像度が 144p，フレームレートが 24fps の Big Buck Bunny を用いる．また，各セグメントの映像時間を 20 秒とする．評価項目は，再生映像における拡大手法の変化回数および平均 LPIPS である．

はじめに，バッチのフレーム数に応じた拡大手法の変化回数を図 7.1 に示す．横軸はバッ

表 7.5 各フレームレートの視聴映像における品質評価

受信映像		提案手法			単純手法			バイキュービック手法		
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Big Buck Bunny	24fps	**28.153**	**0.799**	**0.271**	27,452	0.795	0.279	28.101	0.790	0.320
	30fps	**28.044**	**0.795**	**0.283**	27.377	0.791	0.295	28.001	0.788	0.325
	60fps	**28.013**	**0.786**	**0.314**	27.581	0.783	0.322	27.993	0.783	0.333

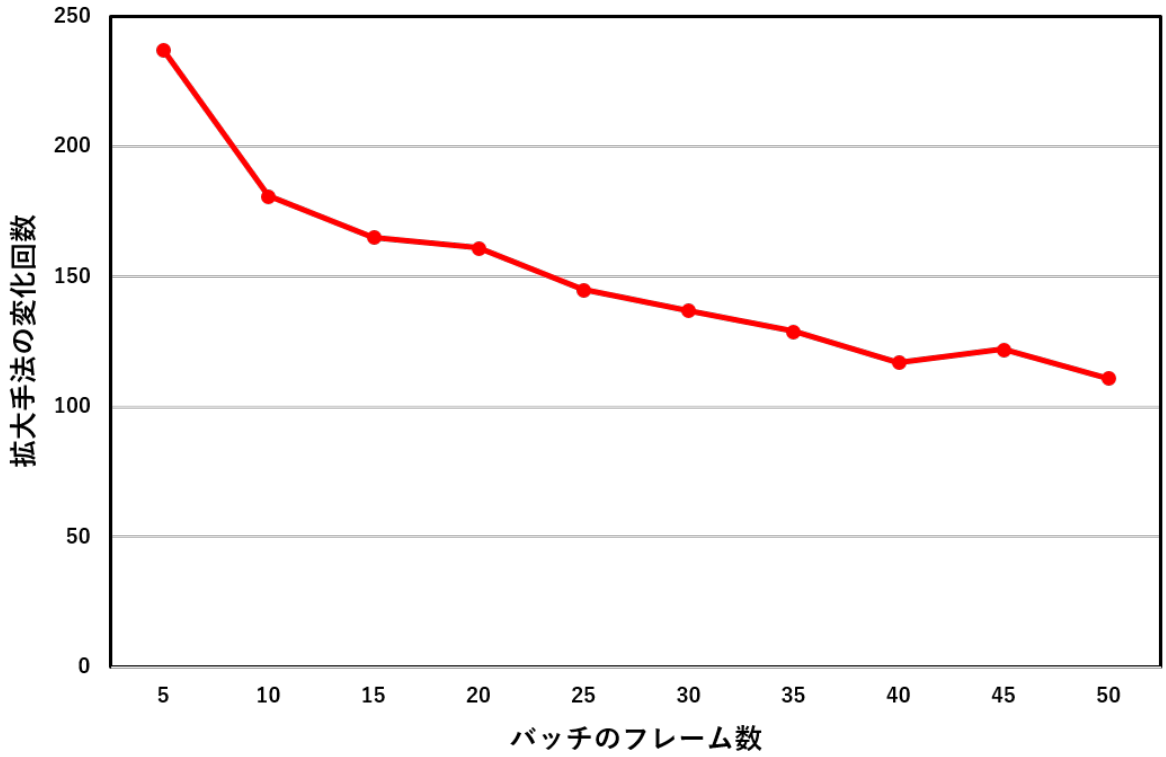

図　7.1　バッチのフレーム数に対する拡大手法の変化回数

チのフレーム数，縦軸は再生映像における拡大手法の変化回数である．図 7.1 より，バッチのフレーム数が大きくなるほど，再生映像における拡大手法の変化回数は少ない．例えば，バッチのフレーム数が 50 の場合，拡大手法の変化回数は 111 回となり，一番少ない．

次に，バッチのフレーム数に応じた平均 LPIPS の評価を図 7.2 に示す．横軸はバッチのフレーム数，縦軸は全フレームの平均 LPIPS である．図 7.2 より，バッチのフレーム数が大きくなるほど平均 LPIPS は大きくなり，映像品質は低下する．例えば，バッチのフレーム数が 5 の場合は平均 LPIPS が 0.2687，バッチのフレーム数が 50 の場合は平均 LPIPS が 0.2701 である．

7.7　セグメントの映像時間による映像品質への影響

提案手法において，セグメントの映像時間の変化に応じた視聴品質を評価する．評価では，解像度が 144p，フレームレートが 24fps の Big Buck Bunny を用いる．また，各バッチのフレーム数は 10 枚とする．評価項目は，平均 LPIPS を用いる．

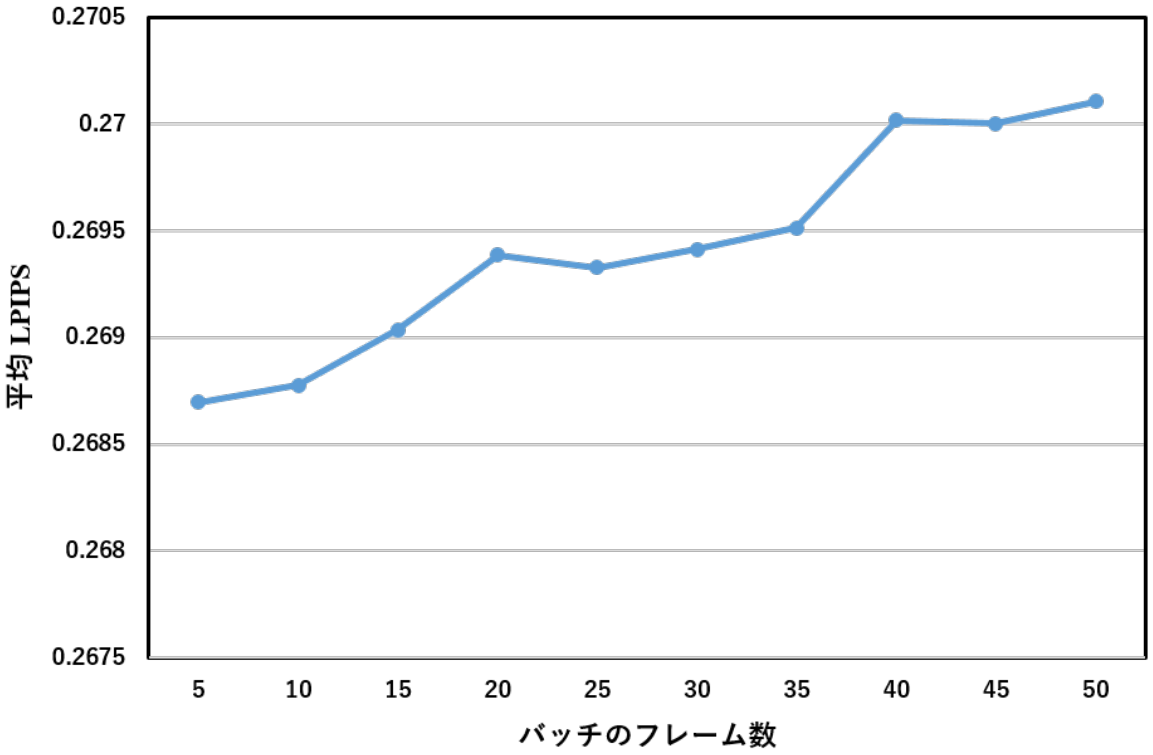

図　7.2　バッチのフレーム数に対する平均 LPIPS

セグメントの映像時間に応じた平均 LPIPS の評価を図 7.3 に示す．横軸はセグメントの映像時間，縦軸は平均 LPIPS である．図 7.3 より，セグメントの映像時間が長くなるほど平均 LPIPS は小さくなり，映像品質は向上する．例えば，セグメントの映像時間が 50 秒の場合，平均 LPIPS は 0.2667 となり，映像品質は最も高い．

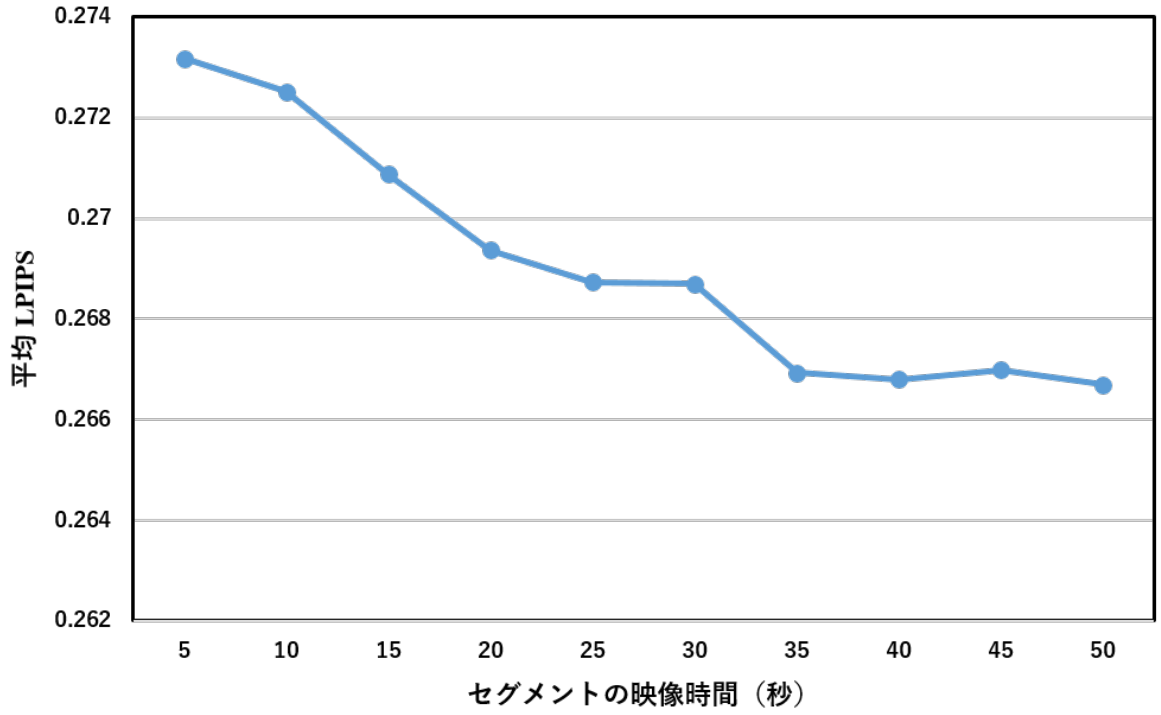

図　7.3　セグメントの映像時間に対する平均 LPIPS

第 8 章

考察

8.1 映像の種類による映像品質への影響

7.3 節の評価結果より，Tears of Steel および Big Buck Bunny では，すべての解像度の映像におけるすべての評価項目において，提案手法が最も高い．提案手法では，特徴量に基づいて超解像フレームを選択するため，他の 2 種類の手法に比べて各フレームの視覚的な平均品質は向上する．

Helzmark Homestead について，180p の映像における SSIM 以外の評価では，単純手法が最も高い．Tears of Steel および Big Buck Bunny は時間的変化が大きい映像である一方で，Helzmark Homestead は時間的変化が小さい映像であり，提案手法による超解像フレームの選択効果は小さい．以上より，提案手法は，時間的変化が少ない映像に対して有用性が低いと考えられる．

バイキュービック手法と単純手法に対して PSNR による評価を比較すると，Tears of Steel における 270p の映像，および Big Buck Bunny における 144p と 180p の映像について，バイキュービック手法が高い．PSNR は，対応するピクセルの画素値を単純に比較して評価する指標であり，バイキュービック法では，周辺画素の平均化によって画素値を補間する．このため，エッジやコーナーが少なく，画素値が近いピクセルが集まるフレームでは，FSRCNN による超解像に比べてバイキュービック法による拡大フレームの PSNR が高くなる場合があるためと考えられる．一方で，バイキュービック手法と提案手法に対して PSNR による評価を比較すると，すべての評価項目においてバイキュービック手法が低い．特徴量が多いフレームを選択したことで，超解像による PSNR の向上が大きいフレームを選択できているためと考えられる．

8.2 映像の解像度による超解像フレーム数への影響

7.4 節の評価結果より，提案手法および単純手法において，映像の解像度が高いほどフレーム 1 枚あたりの超解像処理に必要な時間が長くなる．このため，受信する映像の解像度が高いほど，超解像が行われたフレーム数は少ない．また，提案手法において超解像が行われたフレーム数は，単純手法と比較して，144p の映像では平均して約 609 フレーム，180p の映像では平均して約 225 フレーム，270p の映像では平均して約 100 フレーム少ない．しかし，7.3 節の評価結果より，Tears of Steel および Big Buck Bunny では，すべての解像度の映像におけるすべての評価項目において，提案手法が単純手法に比べて高い．以上より，提案手法では，超解像フレーム数は少ない一方でフレームの平均品質は向上しており，超解像による視覚的な品質向上の効果が高いフレームを選択できている．

8.3 映像のフレームレートによる映像品質への影響

7.5 節の評価結果より，すべてのフレームレートの映像におけるすべての評価項目において，提案手法が最も高い．以上より，フレームレートに関わらず，提案手法が有用である．

PSNR による評価について，提案手法では，フレームレートが高くなると評価が低くなる一方で，単純手法では，24fps や 30fps の映像に比べて 60fps の映像の評価が高い．8.1 節で述べたように，エッジやコーナーが少なく，画素値が近いピクセルが集まるフレームでは，FSRCNN に比べてバイキュービック法による拡大フレームの PSNR が高くなる場合がある．このため，60fps の映像では，FSRCNN でなくバイキュービック法で拡大されるフレームが増加し，単純手法における平均 PSNR が向上したと考えられる．

8.4 バッチのフレーム数による映像の視聴品質への影響

7.6 節の評価結果より，バッチのフレーム数が大きくなるほど，再生映像における拡大手法の変化回数は少ない．提案手法では，バッチ内のフレームは同一の手法で拡大を行うため，バッチのフレーム数を大きくすることで，拡大手法の変化回数を抑えられると考えられる．しかし，バッチのフレーム数が大きくなるほど，平均 LPIPS は大きくなり，各フレームの平均品質は低下する．提案手法では，バッチのフレーム数が大きくなると各セグメントにおけるバッチ数が減少し，少ない数のバッチから超解像を優先するバッチを選択するため，映像品質が低下すると考えられる．

8.5 セグメントの映像時間による映像品質への影響

7.7 節の評価結果より，セグメントの映像時間が長くなるほど平均 LPIPS は小さくなり，映像品質は向上する．提案手法では，セグメントの映像時間が長くなるほど，多くの数のバッチから超解像を優先するバッチを選択できるため，より視覚的な品質向上の効果が高いと予測されるフレームを選択でき，映像品質が向上すると考えられる．

第 9 章

おわりに

本論文では，低品質な映像の受信時に特徴量が多いフレームを優先して超解像処理を行いながら動画を再生する手法を提案した．提案手法では，クライアントは一定時間分の映像をバッファリングしながら再生する．このとき，バッファリング中の映像は，再生が開始されるまでの間で，特徴量が多く視覚的な品質向上の効果が高いと予測されるフレームを優先して超解像処理を行う．評価では，提案手法，特徴量に基づいて超解像フレームを選択しない手法，全てのフレームをバイキュービック法によって拡大する手法の 3 種類を用いて，配信映像に応じた視聴映像の視覚的な品質について，再生フレームの平均 PSNR，平均 SSIM，および平均 LPIPS で比較した．評価の結果，時間的変化が大きい映像の再生時は，解像度やフレームレートに関係なく，提案手法が他の 2 種類の手法と比べて視覚的な品質が高いことを示した．

今後の予定として，提案手法のユーザ評価，他の超解像手法を利用した提案手法の評価，および映像のカットを考慮した超解像フレーム選択手法の提案を行う．

謝辞

本研究を進めるにあたり，ご指導を賜りました後藤佑介准教授に深く感謝いたします.

また，本論文の執筆に当たりまして常に励ましとご協力を頂きました研究室の皆様方に，心より感謝を申し上げます.

参考文献

[1] Cisco Annual Internet Report (2018–2023) White Paper - Cisco (online), available from < https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html > (accessed 2021-01-28).

[2] HTTP Live Streaming: IETF (online), available from < https://tools.ietf.org/html/rfc8216 > (accessed 2021-01-28).

[3] Information Technology - Dynamic Adaptive Streaming over HTTP (DASH) - Part 1: Media Presentation Description and Segment Formats: ISO (online), available from < https://www.iso.org/standard/75485.html > (accessed 2021-01-28).

[4] R. Keys : Cubic Convolution Interpolation for Digital Image Processing, IEEE Trans Acoustic, Speech and Signal Processing, Vol.29, pp.1153-1160 (1981).

[5] C. Dong, C. C. Loy, K. He, and X. Tang : Learning a Deep Convolutional Network for Image Super-Resolution, Computer Vision–ECCV, pp.184–199 (2014).

[6] J. Kim, J. K. Lee, and K. M. Lee : Accurate Image Super-Resolution Using Very Deep Convolutional Networks, IEEE Conference on Computer Vision and Pattern Recognition, pp.1646-1654 (2016).

[7] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang and W. Shi : Photo-realistic single image super-resolution using a generative adversarial network, IEEE Conference on Computer Vision and Pattern Recognition (2017).

[8] M. S. Sajjadi, R. Vemulapalli, and M. Brown : Frame-Recurrent Video Super-Resolution, IEEE Conference on Computer Vision and Pattern Recognition, pp.6626–6634 (2018).

[9] M. Chu, Y. Xie, J. Mayer, L. Leal-Taixé, N.Thürey : Learning Temporal Coherence via SelfSupervision for GAN-based Video Generation, International Conference on Learning Representations (2020).

[10] Z. Zhang and V. Sze : Fast: A Framework to Accelerate Superresolution Processing on Compressed Videos, IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.1015-1024 (2017).

[11] H. Yeo, Y. Jung, J. Kim, J. Shin and D. Han : Neural Adaptive Content-aware Internet Video Delivery, USENIX Symposium on Operating Systems Design and Implementation, pp.645-661 (2018).

[12] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli : Image Quality Assessment: From Error Measurement to Structural Similarity, IEEE Transactions on Image Processing, Vol.13, pp.600-612 (2004).

[13] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik and M. K. Markey : Complex wavelet structural similarity: A new image similarity index, IEEE Transactions on Image Processing, Vol.18, No.11, pp.2385–2401 (2009).

[14] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang : The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, Computer Vision and Pattern Recognition (2018).

[15] E. Rosten and T. Drummond : Machine learning for highspeed corner detection, European Conference on Computer Vision, pp.430-443 (2006).

[16] P. F. Alcantarilla, J. Nuevo and A. Bartoli : Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces, British Machine Vision Conference (2013).

[17] A. Vinay, A. S. Cholin, A. D. Bhat, K. N. B. Murthy and S. Natarajan : An Efficient ORB Based Face Recognition Framework for Human-Robot Interaction, Procedia Computer Science, Vol.133, pp.913–923 (2018).

[18] Y. Li, N. Brasch, Y. Wang, N. Navab and F. Tombari : Structure-SLAM: Low-Drift Monocular SLAM in Indoor Environments, IEEE Robotics and Automation Letters (2020).

[19] T. Hoßfeld, M. Seufert, C. Sieber and T. Zinner : Assessing Effect Sizes of Influence Factors Towards a QoE Model for HTTP Adaptive Streaming, International Workshop on Quality of Multimedia Experience, pp.111–116 (2014).

[20] C. Dong, C. C. Loy and X. Tang : Accelerating the Super-Resolution Convolutional Neural Network, European Conference on Computer Vision, pp.391-407 (2016).

[21] The Apache HTTP Server Project (online), available from
< https://httpd.apache.org/ >
(accessed 2021-01-28).

[22] Tears of Steel — Mango Open Movie Project (online), available from
< https://mango.blender.org/ >
(accessed 2021-01-28).

[23] Big Buck Bunny (online), available from
< https://peach.blender.org/ >
(accessed 2021-01-28).

[24] Herzmark Homestead on Vimeo (online), available from
< https://vimeo.com/226057477/ >
(accessed 2021-01-28).