LI-008

多視点映像符号化における 残差予測を用いた View Synthesis Prediction フレームワーク

View Synthesis Prediction Framework using Residual Prediction for Multiview Video Coding

志水 信哉† 木全 英明† 上倉 一人† 八島 由幸†
Shinya SHIMIZU Hideaki KIMATA Kazuto KAMIKURA Yoshiyuki YASHIMA

1. はじめに

ユーザが実際に撮影を行っているカメラを気にすることなく自由に視点を設定できる自由視点映像や，映像に奥行き感や立体感を付与した立体映像は次世代の映像として高い関心を集めている[1]．どちらの映像も多視点映像を用いて提供することができる．多視点映像は複数台のカメラで撮影した映像群でありそのデータ量は膨大である．そのため，実用化には高効率な符号化が求められる．現在，国際標準化団体 MPEG と ITU-T の合同組織 JVT において MVC(Multi-view Video Coding)の名称で標準化作業が進められている[2][3]．MVC では視点位置を低遅延で切り替えられるなど実用面で重要な機能を持ちつつ高圧縮な符号化を実現することを目標として技術審議が行われている．

多視点映像は同じシーンを複数のカメラで撮影した映像群であるため，あるカメラの映像を別のカメラから効率的に予測することが可能である．同じ時刻の映像を比べた際に生じる視差はカメラ配置とシーンの三次元構造によって変化するため，従来の予測符号化で仮定している並行移動モデルでは映像予測の精度が落ちる．そこで MVC では Image Based Rendering の技術や Computer Vision の技術を使った View Interpolation Prediction や View Synthesis Prediction と呼ばれる新たな映像予測方式の検討が行われている[4]．

我々はこれまで Depth Map と Warping を用いて符号化対象カメラにおける合成映像を生成し，入力映像と合成映像の差分映像を従来の符号化器への入力映像とすることで効率的に多視点映像を符号化するフレームワークを提案してきた[5]．しかし，このフレームワークでは差分映像が入力となるため９ビットの入力/出力信号を扱う符号化/復号器が必要となってしまう．さらに必ず合成映像と差分を取るため，合成映像の品質が悪い部分では符号化効率が悪化してしまうという問題があった．そこで本稿では符号化/復号器へ入力を変更するのではなく，予測信号への操作を行うことで View Synthesis Prediction 誤差に対して残差予測を実現するフレームワークを提案する．

2. View Synthesis/Interpolation Prediction

多視点映像符号化において既に符号化/復号済みの別のカメラで撮影された画像を用いて，符号化/復号対象のカメラにおける画像を合成し，その合成画像を予測画像とする映像予測方式は View Synthesis Prediction と呼ばれている．また２つのカメラの符号化/復号済み画像を用いて行う View Synthesis Prediction を特に View Interpolation Prediction と呼んでいる．これまでに提案されている手法は次の２種る野手法に大別できる．1 つは復号側でシーンの三次元情報を推定し映像合成を行う方法である[6]．この方法では映像合成のために付加情報が必要ないため，ピクセルごとに三次元情報を推定し高い品質の映像を合成し，高い符号化効率を達成することができる．しかし，三次元情報を推定するためには複数の画像が必要となるため，B-View と呼ばれる複数のカメラの画像が参照可能な部分でしか使用することが出来ない．また三次元情報を推定する演算は非常に高負荷なため復号側での演算量増加も深刻な問題である．

もう１つの手法は符号化側でシーンの Depth を推定し，それを符号化して伝送する方式である．この方式では予測効率を高く保ったまま発生する付加情報の符号量を減らすことが課題となる．復号側は送られてきた情報を元に予測信号を生成するため，その演算量は従来の映像符号化とほぼ同じであり，ユーザに高スペックな処理装置を要求しない．本稿で提案するフレームワークもこの種類の手法に属する．

E. Martinian らの方式では，可変サイズの符号化処理ブロック毎に View Synthesis Prediction とその他の予測方式を比較し，View Synthesis Prediction を行う場合にのみブロック単位の Depth を符号化することで発生符号量を最低限に留めている[7]．しかしながらこの手法では符号化対象フレームごとに Depth を符号化するため，異なるフレームで同じ三次元位置を表す情報が符号化されてしまう可能性がある．またブロック毎に Depth を符号化するかしないかを決定するため，Depth の持つ空間的/時間的相関を利用して効率よく符号化することができない．そこで我々はこれまでに多視点映像に対して各時刻で 1 枚 Depth Map を生成し，それをグローバルな Depth 情報として取り扱うことで，Depth 情報を効率的に符号化するフレームワークを提案してきた[5]．グローバルな Depth Map を使用する場合，カメラパラメータの誤差や映像合成時に用いる投影モデルと現実のカメラのミスマッチの影響を予測対象ブロック毎に考慮することが出来ず，合成映像の品質が低下してしまう．この品質低下に対処するために，この方式では入力映像と合成映像の差分映像を予測符号化することで全体として効率的な符号化を実現していた．

しかし，従来のフレームワークには２つの問題点がある．１つは実装コストの問題である．このフレームワークによる MVC 符号化器の構成図全体を図 1，Non-base view の符号化/復号器のブロック図を図 2 に示す．図から分かる通り映像合成による予測は符号化/復号器の外側で行われる前/後処理として扱われる．また，符号化/復号器のブロック構造自体は従来の映像符号化/複合器と全く変わらない．しかしながら入力映像が N ビットの映像信号の場合，合成映像も N ビットの映像信号となるため符号化/復号器の取り扱う信号は N+1 ビットの映像信号となってしまう．一方

†日本電信電話（株）NTT サイバースペース研究所
NTT Cyber Space Laboratories, NTT Corporation

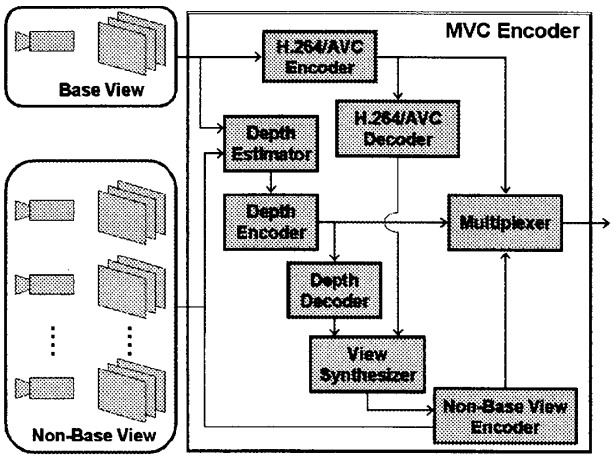

図 1. 想定フレームワーク全体図

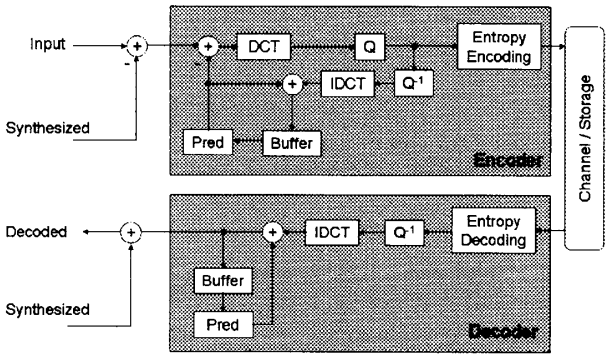

図 2. 従来の Non-Base View Encoder/Decoder

Base view encoder は N ビットの映像信号をそのまま取り扱う．つまり MVC の Encoder/Decoder は 2 つの異なるビット深度の映像信号を取り扱う必要があり実装コストが高くなってしまう．もう 1 つの問題は，必ず入力映像と合成映像との差分映像を生成して符号化するため，映像合成の精度が低い部分では，入力映像をそのまま符号化する場合に比べて符号化効率が低下してしまうことである．本稿ではこれらの問題を解決するためのフレームワークを提案する．

3. 残差予測を用いた View Synthesis Prediction フレームワーク

異なるビット深度の映像信号を取り扱うのを回避するために映像合成による予測を符号化/復号器の内部に含めた構成を提案する．しかしながら E. Martinian らの方式の様に，単純に映像合成で生成された信号そのものを入力映像に対する予測映像とする場合，カメラパラメータの誤差や映像合成時に用いる投影モデルと現実のカメラのミスマッチによる映像合成誤差の影響を取り除くことが出来ない．映像合成誤差の 1 種である合成位置ズレを取り扱うために，実際の予測信号として用いる合成画像上の位置を示すベクトルを符号化する手法があるが，その場合はベクトルを符号化しなくてはならない．また，この方法では平行移動モデルに因らない位置ズレやカメラ間の輝度や色の違いを取り扱うことはできない．

そこで，従来方式のフレームワークで行っていた入力映像と合成映像の差分映像を予測符号化するのと同様の予測を行うために，映像合成予測誤差に対する予測符号化を行う機構を実現する．つまり，映像合成の残差を用いた予測信号に対して映像合成によって生成した信号を加えたものを，入力映像に対する予測信号とする．ここで，映像合成の残差に対する予測を実現するために，バッファにはローカルデコード映像から合成映像を減算した差分映像を蓄積する．この実装の符号化/復号器のブロック図を図 3 に示す．この場合，DCT/IDCT や Q/Q^{-1} で取り扱われる映像信号のビット深度が増加することはない．また，そのビット深度は Base view の映像を符号化する際と同じであるため回路を共有することが可能となる．さらに，入力映像とは異なる N+1 ビットの映像で予測を行うために意味のない予測誤差を符号化して無駄にビットを使用してしまうという従来のフレームワークで回避できない問題を，この構成では予測信号を N ビットにクリッピングすることで回避することも可能となる．

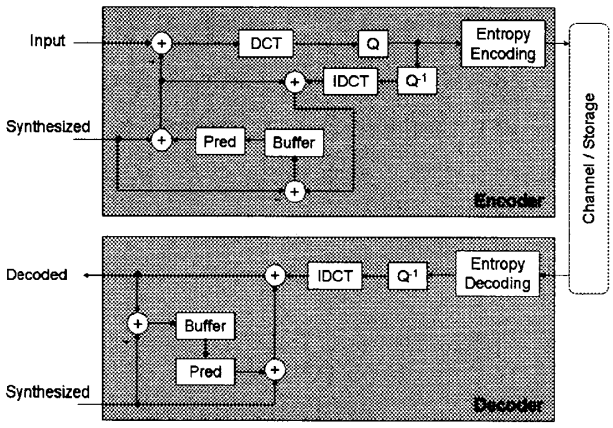

図 3. Residual Prediction Encoder/Decoder

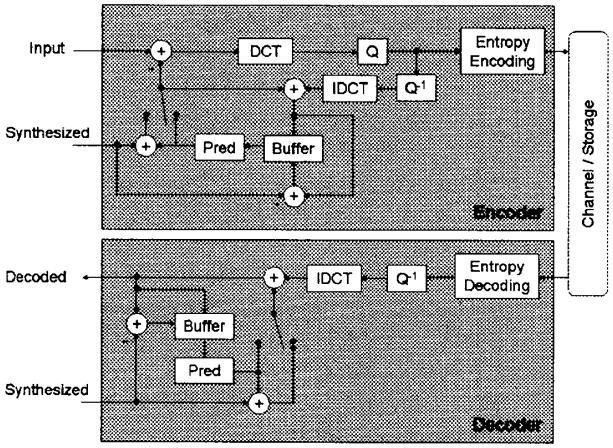

図 4. Adaptive Prediction Encoder/Decoder

合成映像の品質が悪い領域で強制的に映像合成を用いた予測を適用して符号化効率を低下させてしまう問題を解決するために，映像合成を用いた予測を行うか行わないかを適応的に選択可能にする．しかしながら，従来のフレームワークで映像合成による予測信号を使用するかしないかを適応的に切り替えるためには，撮影映像と差分映像との変換を行う前/後処理をスキップさせる必要があるだけでなく，DCT/IDCT や Q/Q^{-1} が扱う映像信号のビット数も変化するため，符号化/復号器のコア部分も切り替える必要がある．このことは Non-base view 単体の符号化/復号器においても

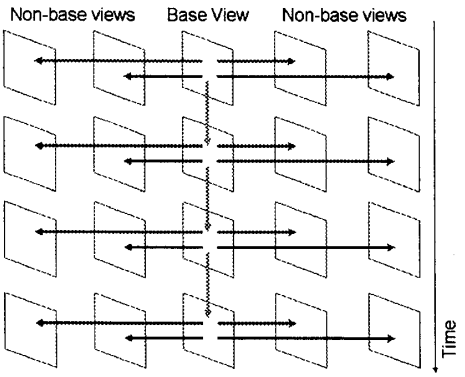

図 5. Prediction Structure for Low Delay

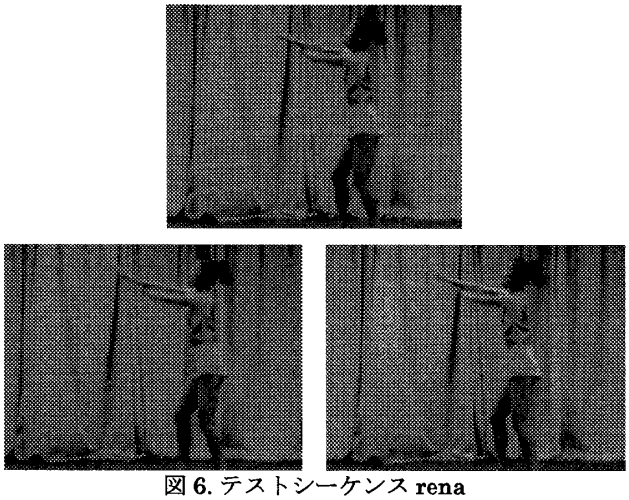

図 6. テストシーケンス rena

複数のビット深度を扱うコアを実装しなくてはならなく回路規模が著しく増加するという問題を生じさせてしまう．

前述の合成映像を予測映像生成のためのオフセットとして用いる提案フレームでは，符号化器への入力は多視点映像の入力映像そのものになり，映像予測もその入力映像に対して行われることになる．つまり，符号化/復号器の内部において合成映像を使用しない一般的な映像予測を行う場合と取り扱う信号のビット信号が統一される．そのため提案フレームワークでは，予測映像生成に通常のローカルデコードを使用し，映像合成による予測の値を加算する部分をスキップするだけで，符号化/復号器のコア部分を共有し少ない回路規模で適応的な予測を行う機構を実現できる．この場合の符号化/復号器のブロック図は図 4 で示される．適応的な処理のために幾つかスイッチが追加され，DPB には差分映像のほかに通常のデコード画像も蓄えられることになる．

4. 実験と考察

4.1. 実験条件

提案するフレームワークによる符号化効率を確認するために，提案手法を標準化の View Synthesis Prediction コア実験で用いられているソフトウェアに実装した[4]．実装では，Intra 予測を View Synthesis Prediction 残差を空間的に予測するモードと置き換え，Inter 予測を動き補償/視差補償/View Synthesis Prediction/時間的 View Synthesis Prediction 残差予測を用いた映像予測モードとした．Inter 予測における予測

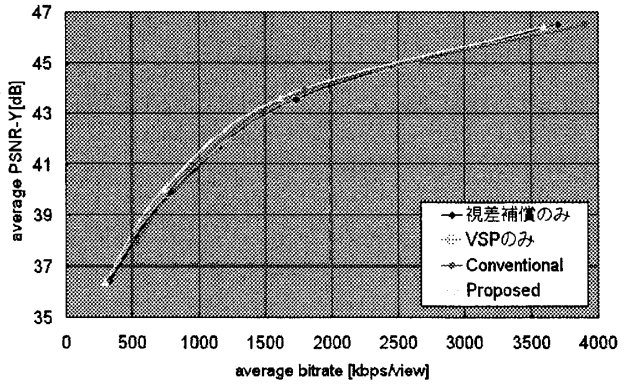

図 7. 実験結果(RD 曲線-全体)

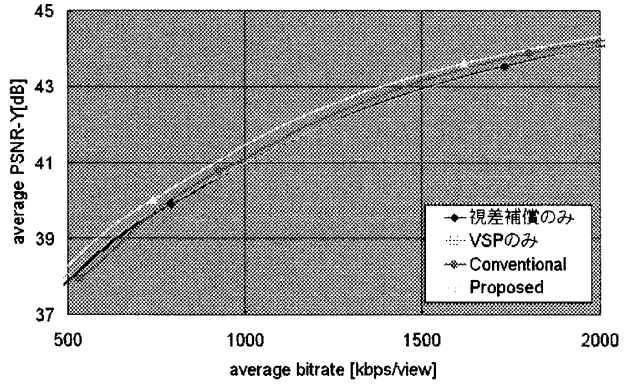

図 8. 実験結果(RD 曲線-拡大)

表 1. 実験結果(Bjontegaard delta)

	BD-bitrate [%]	BD-PSNR [dB]
VSP のみ	-1.07863	0.03201
Conventional	-1.26417	0.07498
Proposed	-7.93795	0.34407

手法の切り替えは参照画像として用いる画像を切り替えることで実現される．

実験では図 5 に示す標準化における実験条件の 1 つである Low Delay 構造を用いた．具体的には Base view を 1 つだけ，多視点映像の中央に位置するカメラに設定し，その他のカメラでは Base view からの view 間の予測のみで符号化する．なお QP22, 27, 32, 37 の固定 QP で符号化を行った．図 6 に実験で用いたテストシーケンス rena におけるカメラの両端と中央の第 1 フレームを示す．

4.2. 実験結果及び考察

視差補償のみ，View Synthesis Prediction（VSP）のみ，従来手法（N+1 ビットの差分映像を符号化），提案手法（VSP + 残差予測）での実験結果を図 7 及び図 8 に示す．図 8 は図 7 の破線で囲まれた部分を拡大したものである．なおこの結果は全カメラの平均であり，視差補償のみ以外の条件では Depth の符号量も含まれる．また視差補償のみを基準とした Bjontegaard delta を表 1 に示す[8]．

View Synthesis Prediction による効率改善は非常に少ないことが分かる．特に中高レート部分ではほぼ視差補償のみと同じ符号化効率であることが分かる．これは，今回用いた Warping による View Synthesis Prediction では，Base view のローカルデコード画像を Warping して合成した画像を参照画像として動き補償が行われるため，視差補償に比べて

ベクトル符号量を抑える働きしか持たないためである．今回実験に用いたシーケンスでは，視差が 0 を中心として凄く限られた大きさでしか存在せず，また視差の持つ空間的な相関のため視差補償ベクトル予測が効果的に働くため効率改善が小さかったと考えられる．視差の範囲が大きく，空間的にも変化が大きなシーケンスの場合，この効果は大きくなると考えられる．しかし，どのようなシーケンスでも高レートになると全体の符号量に対してベクトル符号量の占める割合が少なくなるため効果は非常に少なくなる．

従来手法は中レートにおいて符号化効率を改善できているが，高レートにおいて著しく効率が低下してしまっている．これは高レートでは残差をあまり量子化せずに正確に符号化するため，合成映像の品質が低い領域で大量の残差を符号化していることが影響していると考えられる．

一方，提案手法では効果的に符号化効率を改善できていることが分かる．これは残差予測によって平行移動モデルに因らない位置ズレやカメラ間の輝度や色の違いに対処して View Synthesis Prediction の性能を引き出すことができただけでなく，合成画像の品質に応じて適応的に映像予測手法を選択し，符号化効率が低下するのを抑えることができたためである．品質改善は約 0.34dB であるが，ブロック毎に Depth を符号化する方式での品質改善は参照構造の違いはあるが同じ view 間予測のみが可能なフレームに対して 0.1～0.2dB という報告があるため提案手法は十分有効な手法であると言える[9].

図 7 によると中レートにおいて最も効率がよくなる．原因の 1 つとして今回の実験では提案フレームワークの有効性を調査するために全レートで共通の Depth Map を推定して用いたことが考えられる．つまり使用した Depth Map が中レートにおける量子化で発生する符号化対象の View Synthesis 残差を少なく抑えられていたと考えられる．したがって各レートに対して Depth Map を最適化して求めるか，正確な被写体情報を表す Depth Map を用いることでさらに符号化効率を向上できる可能性がある．

5. おわりに

本稿では，多視点映像符号化において，View Synthesis 画像を予測画像生成のためのオフセットをして用い,View Synthesis 誤差に対して残差予測を行うフレームワークを提案した．これにより従来方式の符号化/復号コアの取り扱う信号のビット深度が増加する課題を解決し実装コスト低減を達成した．またこの実装により実装コストを著しく増加させることなく適応的に予測方式を選択可能な機構を実現することで，合成映像の品質が悪い場合でも強制的に合成映像を用いた予測が行われ符号化が低下してしまう課題も解決した．実験ではこの適応的な予測の効果により提案方式は従来方式に対して，約 7%の符号量削減を達成した．

多視点映像符号化/復号器は同時に複数の映像を取り扱い，また自由視点映像を提供するためにはその映像合成にも高い演算コストがかかるためハードウェア化が重要になる．提案方式による多視点映像符号化/復号器では，コア部分や予測信号生成部などの大部分が H.264/AVC で使用されているコンポーネントと全く同じである．そのため H.264/AVC 符号化/復号器の実装を流用することができ，ハードウェア化の開発コストを低く抑えることが可能である．これは三次元映像配信サービスの発展を考えた際に非常に大きな利点となると考えられる．

今後の課題として，今回はテストしなかった View Synthesis Prediction 残差の時間的予測の性能評価が挙げられる．従来方式で差分映像を動き補償しながら符号化する方式は非常に効果的であったため，時間的な残差予測を導入することでさらに高い符号化効率を達成できると考えられる．更に，今回の実装では Intra 予測のシンタックスを空間的残差予測に割り当ててしているため，オクルージョン部分などの Intra 予測が効果的な部分で符号化効率が低下してしまっている．提案機構では従来の映像予測方式を取り扱うことが可能なため Intra 予測と空間的残差予測に適切なシンタックスを割り当てて，適応的に切り替える仕組みを実現することができる．これによって更なる符号化効率改善が見込めるだけでなく，オクルージョンが多く発生しているような多視点映像に対しても提案方式で効率的な符号化を実現できるようになると考えられる．

また，今回用いた Depth Map はターゲットレートに関係なく，16x16 ブロック毎に 1 つの Depth を求めたものを用い，その符号化にはロスレス符号化を適用した．この方式は必ずしも最適なものではない．Depth 推定の方法だけでなく，提案フレームワークに適した Depth 解像度や符号化手法を検討することも今後の課題である．

謝辞

本研究で用いたテストシーケンスを提供して頂いた名古屋大学工学研究科の谷本正幸教授ならびに谷本研究室に深く感謝申し上げます．

参考文献

[1] A. Smolic, and P. Kauff, “Interactive 3-D Video Representation and Coding Technologies,” *In Proc. The IEEE, Special Issue on Advances in Video Coding and Delivery*, vol. 93, no. 1, pp. 98-110, Jan. 2005.

[2] 木全英明, “MPEG 3DAV 国際標準化の動向,” *映像情報メディア学会誌*, vol.60, no.2, pp.143-149, Feb. 2006.

[3] 木全英明, “多視点映像符号化 MVC の国際標準化動向,” *映像情報メディア学会誌*, vol.61, no.4, pp.426-430, Apr. 2007.

[4] H. Kimata, “CE6: View Interpolation Prediction for MVC”, *Doc. JVT-V306*, Marrakech, Morocco, Jan. 2007.

[5] S. Shimizu, M. Kitahara, K. Kamikura, and Y. Yashima, “Multi-view video coding based on 3-D warping with depth map,” *In Proc. PCS2006*, Apr. 2006.

[6] K. Yamamoto, T. Yendo, T. Fuji, M. Tanimoto, M. Kitahara, H. Kimata, S. Shimizu, K. Kamikura, and Y. Yashima, “Multi-view Video Coding using View-interpolated Reference Images,” *In Proc. PCS2006*, Apr. 2006.

[7] E. Martinian, A. Behrens, J. Xin, and A. Vetro, “View Synthesis for Multiview Video Compression,” *In Proc. PCS2006*, Apr. 2006.

[8] G. Bjontegaard, “Calculation of average PSNR differences between RD-curves,” *VCEG Contribution VCEG-M33*, Austin, Apr. 2001.

[9] S. Yea, and A. Vetro, “Report of CE6 on View Synthesis Prediction,” *Doc. JVT-W059*, San Jose, USA, Apr. 2007.