Hybrid Learning-based and HEVC-based Coding of Light Fields

Milan Stepanov, Giuseppe Valenzise, Frédéric Dufaux

HYBRID LEARNING-BASED AND HEVC-BASED CODING OF LIGHT FIELDS

Milan Stepanov, Giuseppe Valenzise, Frédéric Dufaux

Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes
91190, Gif-sur-Yvette, France

ABSTRACT

Light fields have additional storage requirements compared to conventional image and video signals, and demand therefore an efficient representation. In order to improve coding efficiency, in this work we propose a hybrid coding scheme which combines a learning-based compression approach with a traditional video coding scheme. Their integration offers great gains at low/mid bitrates thanks to the efficient representation of the learning-based approach and is competitive at high bitrates compared to standard tools thanks to the encoding of the residual signal. The proposed approach achieves on average 38% and 31% BD rate saving compared to HEVC and JPEG Pleno transform-based codec, respectively.

Index Terms— light field, deep learning, HEVC, JPEG Pleno

1. INTRODUCTION

Light fields (LFs) capture light intensity of scene objects from various angles. The captured information offers novel applications such as refocusing and a perspective change at the cost of increased dimensionality and thus storage demand. The necessity of efficient compression methods was acknowledged by JPEG committee which started the JPEG Pleno initiative in order to provide a standard framework for the representation and coding of plenoptic data. In particular, some coding tools are proposed and optimized to encode 4D light fields [1].

On the other side, some recent compression approaches based on learning focus on optimizing the whole coding pipeline in an end-to-end fashion [2]. Typically, these schemes employ deep auto-encoders to learn good representations of data, which are then quantized and entropy coded. The potential of these methods has been shown by their competitive performance in terms of objective and subjective metrics compared to traditional coding schemes [3]. However, it has been found that auto-encoder based approaches provide significant gains at low bitrates, while they generally fail to provide high quality and near-lossless reconstructions at high bitrates. This phenomenon is mainly due to the nature of autoencoders, which are intrinsically lossy. Conversely, traditional codecs are designed to span the full quality range, and in particular to provide near lossless performance at higher bitrates.

In order to incorporate the benefits of both approaches, we aim at overcoming the observed lack of scalability of deep learning approaches by adding an enhancement layer. We propose a hybrid scheme consisting of a base layer which provides high gains at low/mid bitrates and serves as an efficient predictor for high bitrates. This is complemented by an enhancement layer which allows the coding of the residual signal via a traditional coding scheme and provides improved performance at high bitrates. Furthermore, we explore various traditional coding schemes for the residual signal and show that even with a simple approach such as scalar quantization it is possible to achieve significant gains with respect to the base layer and to be competitive with state-of-the-art LF codecs.

2. RELATED WORK

We divide the related work into two categories: the LF compression using traditional approaches based on conventional prediction and transformation pipelines, and end-to-end learning-based image compression methods.

High Efficiency Video Coding (HEVC) and its extensions significantly influenced LF coding [4][5][6]. Typically, HEVC is used to encode a set of key views which is at the decoder used to reconstruct the complete LF. Zhao et al. [7] reconstruct the LF using a linear combination of decoded views. Jiang et al. [8] recovers missing views via warping and inpainting of occluded regions. Viola et al. [9] use graph learning method to learn disparities between views and restore LF. These approaches are based on schemes including different functional blocks and therefore difficult to optimize together. Conversely, an end-to-end scheme learns a single function which jointly optimizes and integrates all needed operations. JPEG Pleno provided a coding tool for the light field modality which works in two modes: the prediction (WaSP) and the transformation (MuLE) [10]. The prediction mode uses disparity-based warping and view merging to predict views, while the transformation mode exploits the redundancy in LF blocks using 4D Discrete Cosine Transform. Although, the solutions are tailored for LF compression, they are limited in the same way as the previous approaches.

End-to-end learning-based compression has been recently proposed in [11] [12] [13] [14] and has gained huge popularity due to its ability to replace the whole traditional compression pipeline with a single function. Ballé et al. proposed an end-to-end compression approach which consists of analysis and synthesis functions corresponding to an encoder and a decoder in conventional pipelines, respectively, plus a uniform quantizer. In addition, they propose a differentiable quantization mechanism allowing to optimize the rate-distortion (RD) function directly [11]. Theis et al. [13] propose a similar approach but deals with quantization and bitrate estimation in different manner. Rippel et al. [14] propose a real-time codec which applies a pyramidal analysis for the feature extraction and an adaptive coding module and regularization. Conversely to the previous approaches, Toderici et al. [12] overcomes the necessity to train a separate model for each lambda value in the RD function by adopting the encoding in a progressive manner.

Our approach extends the work of Ballé et al. [11], and operates on data of a higher dimension, requiring a careful design of the network architecture to handle a particular filtering across different views. Furthermore, motivated by the limitations of auto-encoders in providing high-quality reconstructions, we also introduce an enhancement layer to encode the residual signal.

3. PROPOSED METHOD

Fig. 1 shows our proposed scheme. It consists of a base layer illustrated with red blocks and an enhancement layer denoted by cyan blocks. The layers are presented in detail in the following subsections.

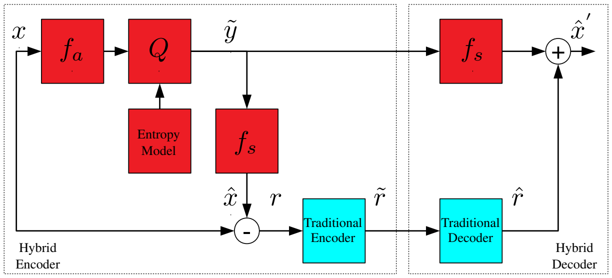

Fig. 1. Proposed coding scheme.

3.1. Base Layer

As a base layer, we propose an end-to-end trained compression scheme based on the recent work of Ballé et al. [11] with added modifications to adjust to the LF structure. More specifically, the scheme takes as input a LF image, reshapes it by extracting sub-aperture views and stacking the views along the third dimension following horizontal raster scan order, i.e. row by row selection. The compression scheme is comprised of three functional blocks: an analysis function $f_a(\theta)$ which creates a more compact representation of the input $y = f_a(x)$, a quantization block or an entropy bottleneck $Q(\eta)$, which provides quantized version of y, $\tilde{y}$, and a synthesis function $f_s(\phi)$ whose goal is to reconstruct the input from the quantized compact representation $\hat{x} = f_s(\tilde{y})$.

The analysis function comprises a set of sequential nonlinear, downsampling, and convolutional layers while the synthesis function is a symmetric counterpart of the encoding function with downsampling layers replaced by upsampling layers. More specifically, each layer utilizes 2D filters, which slide along spatial dimensions and accumulate the contribution of each feature map in the data signal in order to jointly learn structures in spatial and angular domains. The entropy bottleneck works in two modes depending on the inference phase, i.e. the training or testing phase. During the training phase the entropy bottleneck adds a uniform noise $U(-0.5, 0.5)$ to the transformed representation of the input to approximate quantization in a differentiable manner. Furthermore, the entropy bottleneck learns the probability density function of each feature map of the compacted representation and utilizes them for entropy coding. A detailed description of the network parameters is summarized in Fig. 2.

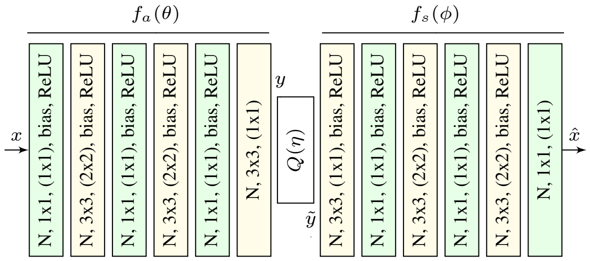
$f_a(\theta)$
$f_s(\phi)$
x
N, 1x1, (1x1), bias, ReLU
N, 3x3, (2x2), bias, ReLU
N, 1x1, (1x1), bias, ReLU
N, 3x3, (2x2), bias, ReLU
N, 1x1, (1x1), bias, ReLU
N, 3x3, (1x1)
y
$Q(\eta)$
$\tilde{y}$
N, 3x3, (1x1), bias, ReLU
N, 1x1, (1x1), bias, ReLU
N, 3x3, (2x2), bias, ReLU
N, 1x1, (1x1), bias, ReLU
N, 3x3, (2x2), bias, ReLU
N, 1x1, (1x1)
$\hat{x}$

Fig. 2. Neural network architecture. The parameters in each block denote the number of filters, spatial extent of the filter, stride, the usage of the bias and the activation function.

Weights of the analysis function, the entropy bottleneck and the synthesis function, θ, η and ϕ respectively, are learned by minimizing the RD function $J(\theta, \eta, \phi; x) = R(\tilde{y}) + \lambda D(x, \hat{x})$ where the rate $R(\tilde{y})$ is modeled with the entropy of the compressed bottleneck, the distortion D is the mean square error between the input x and the decoded LF $\hat{x}$, and the parameter λ governs the trade-off between the rate and the distortion. We trained five models by selecting five different lambda values. For each model, the weights are learned using Adam optimizer with a learning rate of 1×10^{-4} and 1×10^{-3}, for θ and ϕ, and η respectively. The final bitstream is obtained by packing encoded coefficients, the LF size and the lambda parameter.

Bikes (I01) Danger de Mort (I02) Stone Pillars Outside (I04) Fountain and Vincent 2 (I09)

Fig. 3. Test images for the quality evaluation.

3.2. Enhancement layer

We observe that the auto-encoder based approach used in the base layer reaches a saturation in performance at higher bitrates. A possible solution to this problem is increasing the capacity of the network. However, this requires to use increasing model complexity when approaching higher quality points. Instead, we propose to introduce an enhancement layer to encode the residual signal between the original light field and the reconstruction from the base layer. The advantage of this hybrid approach is that the residual coding allows incorporating any traditional image coding scheme. We compute the residual signal by subtracting an input LF and its prediction obtained using the base scheme followed by cropping to a fixed range of values and translation to positive values. For the purpose of an enhancement layer, we explore three coding schemes. We evaluate scalar quantization followed by sample entropy estimation due to its simplicity. Furthermore, we evaluate HEVC coder using Intra and Inter prediction to capture the remaining correlation in spatial and angular domains.

4. RESULTS

4.1. Testing conditions

In our test, we follow test conditions proposed by JPEG Pleno (CTC) [1] with slight modifications due to the setup of our approach. Namely, as our approach operates on the luminance component only, we convert LF to YCbCr colour space and set chroma components to neutral colour. We use the EPFL LF image dataset [15] and select 4 test images as shown in Fig. 3 and divide the rest of the images into training and validation sets in the ratio 80/20. Each raw image is decoded using LFToolbox version 0.4 [16] [17], and a subset of 13 × 13 sub-aperture views, each having the spatial resolution of 625 × 434 pixels is selected to generate a LF image. The proposed approach is compared with HEVC (x265 implementation as defined in the JPEG Pleno CTC) and JPEG Pleno Verification Model 2.0 (VM 2.0) in terms of average PSNR and SSIM. Prior to x265 encoding, a LF is converted to YCbCr 4:4:4 10-bit colour space and arranged in a pseudo-video sequence following serpentine scanning order as defined in CTC. We evaluate VM 2.0 in the transformation mode (MuLE) as it is more efficient for the coding of lenselet data. Four bitrates are used to generate RD curve: 0.75 bpp (bits per pixel), 0.1 bpp, 0.02 bpp and 0.005 bpp.

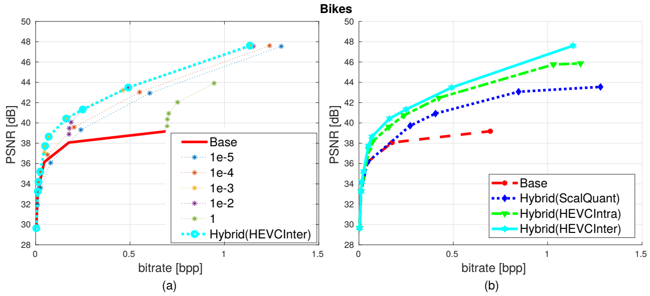

Fig. 4. RD curve of the base layer, local RD curves obtained for each lambda value and the RD curve obtained by computing the convex hull around all the point (a) and the comparison of the base layer and the improved versions (b) for the LF image *Bikes*.

4.1.1. *Joint rate-distortion curve*

The contribution of the enhancement layer is evaluated in terms of RD by generating a RD curve that corresponds to the joint coding of the base layer and the enhancement layer. The residual image is encoded using the proposed approaches for different values of the quantization parameter providing local RD curves around each RD point of the base layer. Fig. 4 (a) illustrates the RD curve of the base layer (red) and the points of local RD curves (asterisks). The final RD curve is obtained by computing the convex hull around all the points (cyan).

4.2. The comparison of enhancement layers

	Base		
Im.	ScalQuant	HEVCIntra	HEVCInter
I01	-13.2317%	-27.4263%	**-31.9636%**
I02	-4.9793%	-13.8663%	**-16.0852%**
I04	-8.3554%	-22.1310%	**-22.6794%**
I09	-14.4142%	-22.2486%	**-31.2664%**

Table 1. BD rate savings of the proposed hybrid approach with three types of enhancement layer against the base layer.

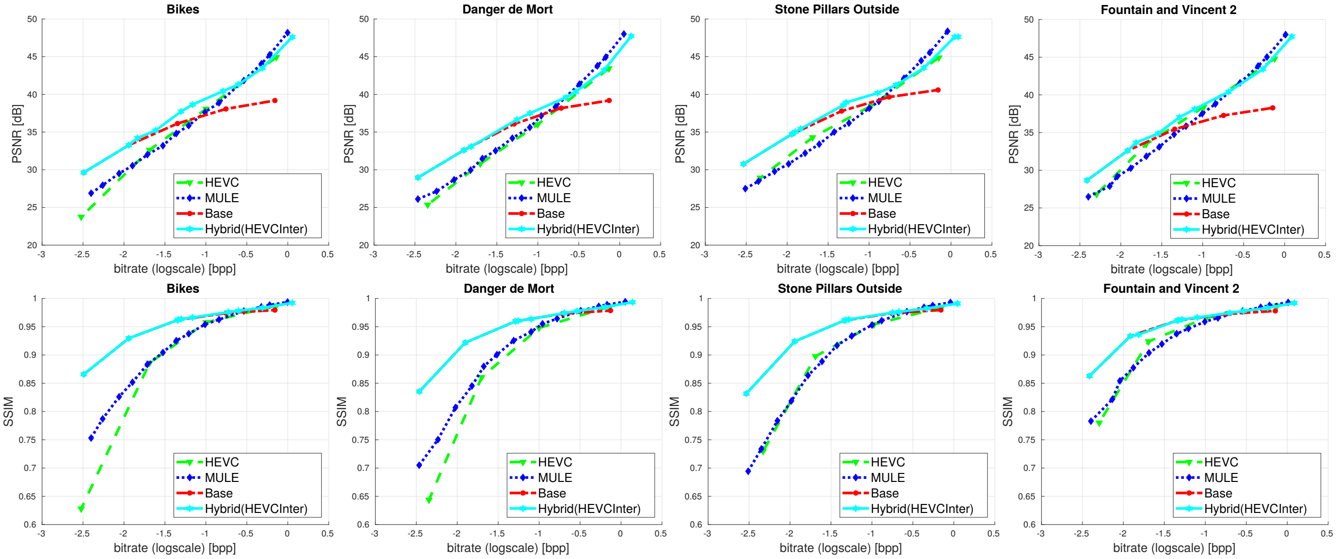

Fig. 5. Performance with respect to PSNR (upper row) and SSIM (bottom row): proposed Hybrid approach using HEVC Inter for the enhancement layer, learning-based base layer, HEVC, MuLE.

Fig. 4 (b) compares the performances of the base layer against the proposed extensions for the LF *Bikes*. It can be noticed that at low bitrates enhancement layers do not improve performance suggesting the superiority of the base layer. On the contrary, at high bitrates we can notice the benefit of adding the enhancement layer and note that even a simple method such as scalar quantization provides $\sim 5-15\%$ savings comparing to the base layer. Further improvements are obtained with HEVC's Intra and Inter prediction modes which gain by exploiting the spatial and inter-view correlations in the residual signal. The quantification of the performance is presented in Table 1 for all test contents.

4.3. The comparison with respect to anchors

We compare the best performing proposed hybrid approach with the original base layer and two anchor methods proposed by JPEG Pleno, and provide RD comparison in terms of PSNR and SSIM in Fig. 5 and BD rate in terms of PSNR in Table 2. The hybrid approach gains at lower bitrates thanks to the learning-based base layer, while the introduction of the enhancement layer increases the performance and makes the approach competitive with the anchors at high bitrates. Nevertheless, overall performance suggests significant gains of the approach against the two anchors ranging from $\sim$ 16% to $\sim$ 50% saving with respect to HEVC and from $\sim$ 25% to $\sim$ 40% saving with respect to MuLE.

5. CONCLUSION

We propose a hybrid coding scheme for LF based on a learning-based approach and HEVC in order to overcome the saturation in performance at high bitrates of auto-encoders. Furthermore, we show that the proposed approach achieves better performances against state-of-the-art anchors. More precisely, the results show that it is possible to greatly improve the performance at high bitrates and moderately at mid bitrates compared to the learning-based base approach. At the same time, the latter provides highly competitive gains at low bitrates, demonstrating the superiority of learning-based representations also for the case of LF.

In future work we aim to optimize the base layer and the enhancement block jointly in an end-to-end fashion and to add the coding of chroma components.

	Hybrid(HEVCInter)		
Im.	Base	MuLE	HEVC
I01	-31.9636%	-36.0352%	-40.3208%
I02	-16.0852%	-25.1350%	-49.4795%
I04	-22.6794%	-39.4541%	-45.7385%
I09	-31.2664%	-25.0123%	-16.3930%
Avg.	-25.4987%	-31.4092%	-37.9829%

Table 2. BD rate savings of the best proposed Hybrid approach using HEVC Inter for the enhancement layer, against three anchors: learning-based base layer, MuLE, HEVC.

6. ACKNOWLEDGEMENT

This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 765911 (RealVision).

7. REFERENCES

[1] ISO/IEC JTC 1/SC29/WG1 N81022, "JPEG PLENO - light field coding common test conditions," 2018.

[2] Giuseppe Valenzise, Andrei Purica, Vedad Hulusic, and Marco Cagnazzo, "Quality assessment of deep-learning-based image compression," in *International Workshop on Multimedia Signal Processing (MMSP)*, 2018.

[3] ISO/IEC JTC 1/SC29/WG1 N85013, "Performance evaluation of learning based image coding solutions and quality metrics," 2019.

[4] Alexandre Vieira, Helder Duarte, Cristian Perra, Luis Tavora, and Pedro Assuncao, "Data formats for high efficiency coding of lytro-illum light fields," in *International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2015.

[5] Waqas Ahmad, Roger Olsson, and Marten Sjostrom, "Interpreting plenoptic images as multi-view sequences for improved compression," in *International Conference on Image Processing (ICIP)*, 2017.

[6] A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, and F. Dufaux, "Integral images compression scheme based on view extraction," in *European Signal Processing Conference (EUSIPCO)*, 2015.

[7] Shengyang Zhao and Zhibo Chen, "Light field image coding via linear approximation prior," in *International Conference on Image Processing (ICIP)*, 2017.

[8] Xiaoran Jiang, Mikael Le Pendu, and Christine Guillemot, "Light field compression using depth image based view synthesis," in *International Conference on Multimedia & Expo Workshops (ICMEW)*, 2017.

[9] Irene Viola, Hermina Petric Maretic, Pascal Frossard, and Touradj Ebrahimi, "A graph learning approach for light field image compression," in *Applications of Digital Image Processing XLI*, 2018.

[10] Peter Schelkens, Pekka Astola, Eduardo A. B. Da Silva, Carla Pagliari, Cristian Perra, Ioan Tabus, and Osamu Watanabe, "JPEG pleno light field coding technologies," in *Applications of Digital Image Processing XLII*, 2019.

[11] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.

[12] George Toderici, Sean M. O'Malley, Sung Jin Hwang, Damien Vincent, David Minnen, Shumeet Baluja, Michele Covell, and Rahul Sukthankar, "Variable rate image compression with recurrent neural networks," in *International Conference on Learning Representations*, 2016.

[13] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár, "Lossy image compression with compressive autoencoders," in *International Conference on Learning Representations*, 2017.

[14] Oren Rippel and Lubomir Bourdev, "Real-time adaptive image compression," in *International Conference on Machine Learning*, 2017.

[15] Martin Rerabek and Touradj Ebrahimi, "New light field image dataset," in *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[16] Donald G. Dansereau, Oscar Pizarro, and Stefan B. Williams, "Decoding, Calibration and Rectification for Lenselet-Based Plenoptic Cameras," in *Conference on Computer Vision and Pattern Recognition*, 2013.

[17] Donald G. Dansereau, Oscar Pizarro, and Stefan B. Williams, "Linear volumetric focus for light field cameras," *ACM Transactions on Graphics*, 2015.