Optimized Spatial Recurrent Network for Intra Prediction in Video Coding

Yueyu Hu, Wenhan Yang, Sifeng Xia, Jiaying Liu*
Institute of Computer Science and Technology, Peking University, Beijing, P. R. China, 100871

***Abstract*—Intra prediction in modern video codecs is able to efficiently reduce spatial redundancy in video frames. With preceding pixels as context, traditional intra prediction schemes generate linear predictions based on several predefined directions (*i.e.* modes) for the current prediction unit (PU). However, these modes are relatively simple and are not able to handle complex textures, which leads to additional bits encoding the residue. In this paper, we design a convolutional neural network (CNN) guided spatial recurrent neural network (RNN) to improve the intra prediction in High-Efficiency Video Coding (HEVC). By exploring the correlations between pixels, the network learns to generate prediction signal in a progressive manner. The progressive model solves the problem of asymmetry in intra prediction naturally. As the model is designed for global context modeling, no flags for intra prediction modes selection need to be encoded. Our proposed intra prediction scheme achieves on average 1.2% bit-rate saving compared with HEVC.**

***Index Terms*—Video Coding, Intra Prediction, Recurrent Neural Network, HEVC**

I. Introduction

Intra prediction is a basic component of modern video codecs (e.g. HEVC). It significantly saves bit-rate by reducing spatial redundancy. HEVC uses up to 35 modes for directional intra prediction [1]. In the rate-distortion optimization (RDO) [2] scheme of HEVC, the codec searches for the best prediction result among the 35 modes including DC, planar and 33 directional modes, and comes out with the most appropriate mode.

However, the prediction performance of HEVC can be further improved. Firstly, for the single-line reference scheme in HEVC, prediction signals are generated in accordance with the most adjacent line of the available reconstructed blocks. As a consequence, in low bit-rate configurations, the predictions tend to be inaccurate. Secondly, directional intra prediction cannot handle complex texture. To address the drawback of the single-line reference scheme, a multi-line scheme is introduced in [3, 4]. By expanding the reference area to more reference lines, interference of noises produced by aggressive quantization is reduced. However, these methods only expand the reference area, the structural correlations of pixels in the reference area are not explored. Thus, the improvement of the coding performance is limited.

Recently deep learning methods emerge for image and video compression and processing tasks. Involving deep models in video coding has been initially studied in recent years [5–8]. Deep neural networks can automatically learn the end-to-end mapping of inputs and outputs. It can also be easily accelerated using large-scale parallel programming. In [9], CNN has been utilized for mode decision as it has a strong potential of capturing global feature from image data. In [7, 10] fully-connected (FC) network and CNN are exploited directly in intra prediction. By training a network to build a mapping from the reference samples to the prediction signal, FC networks and CNN show improvement in rate-distortion performance compared with HEVC.

However, previously proposed deep-network-based intra prediction methods have drawbacks. For CNN based methods, the network is not capable of handling asymmetric image completion tasks, as the whole input block is convolved unconditionally. Large areas with no texture information interferers the extraction of spatial features. For FC networks, the correlations among pixels in the reference area are neglected. As a consequence, it is hard for FC networks to generate predictions for sharp edges.

To address the deficiencies of the previous model discussed above, in this paper, we propose a CNN guided spatial recurrent neural network (RNN) to improve the intra prediction in HEVC. By exploring the correlations between pixels, the network learns to generate prediction signal in a progressive manner. As no information from the unknown regions is used before the prediction process, the model solves the problem of asymmetry naturally. This network achieves promising performance gain compared with HEVC.

The rest of the paper is organized as follows. In Section II, we analyze the different reference scheme for traditional and deep learning based intra prediction. In Section III, the architecture of spatial RNN for pixel modeling is explained. In Section IV, we demonstrate the experimental results of our network. In Section V we draw a conclusion.

II. Blocking Reference Scheme

HEVC uses extended reference blocks and richer directional prediction modes compared with its predecessor AVC/H.264 [11] to provide better intra prediction accuracy and enhance overall coding performance. In HEVC intra prediction, a PU with size $M \times M$ will be provided with $4M + 1$ reference samples, as is shown in Fig. 1. Although the reference area is expanded, only the adjacent lines of the available coded blocks are used as the reference. Thus, noise in the original frame or generated during the quantization results in false textures in the reference line. The impact of

* Corresponding author. This work was supported by National Natural Science Foundation of China under contract No. 61772043.

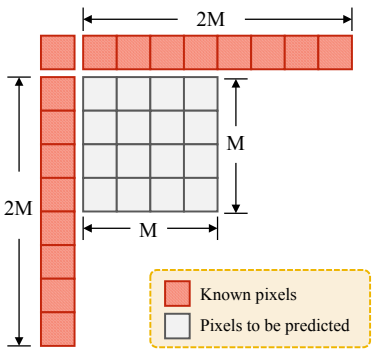

Fig. 1. The Γ-like reference context in HEVC. Reconstructed pixels from the left below and the right above will be included in the reference samples.

the noise grows as the quantization parameter increases. One way to address this issue is to expand the area of reference samples. In a multi-line reference scheme, multiple preceding lines are jointly exploited for reference of predicton, as is shown in Fig. 2 (a). However, the expanding reference area brings additional complexity to directional intra prediction. The improvement of the expansion reaches a limit for such prediction methods.

To further utilized the preceding pixels, our proposed method adopts a block-level reference scheme, where at most five available encoded blocks can be utilized, as shown in Fig. 2 (b). As the proposed method is able to handle spatial correlations, the reference pixels can be viewed as patches other than lines during the generation of predictions.

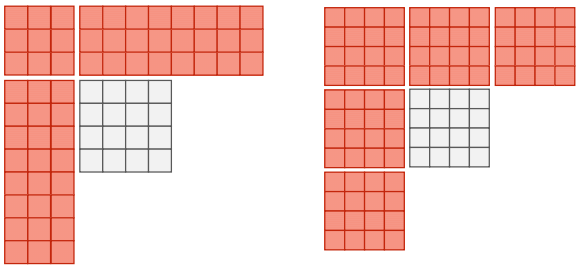

Fig. 2. Different reference scheme to enable robust referencing and reduce the influence of noise. (a) Multi-line reference scheme. (b) Block-Level reference scheme.

III. Spatial RNN for Intra Prediction

A challenge for designing a neural network for intra prediction lies in the asymmetry of the inputs. In this problem, the reference samples are distributed in the blocks on the left and the above. A large area of unknown pixels on the right below of the context provides no information for prediction process. This asymmetric property causes two main problems. First, the network needs to be deep to provide a large enough receptive field for each neuron. Such a deep network is hard to train. Second, the area with no information is convolved unconditionally. As illustrated in Fig. 3, the receptive field for a right below target pixel will be largely covered by the unknown area. It interferes the training of filters and results in inaccurate predictions.

One way to deal with the problem is to use FC neural network. In an FC network, the inputs are connected to every input dimension with densely connected neurons, allowing the network to do a global optimization. However, local features of spatially distributed pixels are not effectively extracted in an FC network. It it hard for such a network to generate sharp edges.

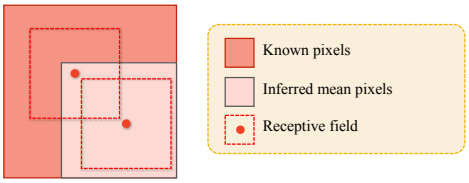

Fig. 3. Receptive field of a right-below point can be largely covered by filled data. Points in the right-bottom area are generally mapped from previously filled singals.

To address the issues mentioned above, we propose to exploit spatial RNN for intra prediction. The architecture of the network is inspired by [12]. As is illustrated in Fig. 4, the network consists of two parts. We first map the input pixels to feature space with convolutional layers. As these layers are shallow and the sizes of the kernels are relatively small, global interference of the large missing area is not significant.

Before the prediction, the feature maps are re-sampled to several scales, making the network compatible for variable content scale in videos. After the concatenation of each scales, the network progressively generates predictions for the feature maps. We define the feature map as $\mathbf{X}$. It is viewed as a stack of horizontal planes $\mathbf{X}^h = \{\mathbf{X}_0^h, \mathbf{X}_1^h, \ldots, \mathbf{X}_{n-1}^h\}$ or a stack of vertical planes $\mathbf{X}^v = \{\mathbf{X}_0^v, \mathbf{X}_1^v, \ldots, \mathbf{X}_{n-1}^v\}$, where each element in the stack represents a feature vector. We take the vertical generation process as an example. Under the assumption that the distributions of local features are continuous, this process is formulated as,

$$\tilde{\mathbf{X}}_i^v = \mathcal{F}\left(\mathbf{X}_i^v, \mathbf{X}_{i-1}^v, \theta_v\right), \tag{1}$$

where $\tilde{\mathbf{X}}_i^v$ is the predicted feature vector for the i^{th} plane and $\mathcal{F}$ is the function to generate the prediction signals from previous observations together with the current input feature vector.

Our network automatically learns the mapping function $\mathcal{F}$ using spatial RNN with Gated Recurrent Units (GRU). A traditional GRU can be formulated as follows,

$$\begin{aligned} \mathbf{z}_t =& \sigma(\mathbf{W}^{\mathbf{z}}\mathbf{x}_t + \mathbf{U}^{\mathbf{z}}\mathbf{h}_{t-1}), \\ \mathbf{r}_t =& \sigma(\mathbf{W}^{\mathbf{r}}\mathbf{x}_t + \mathbf{U}^{\mathbf{r}}\mathbf{h}_{t-1}), \\ \mathbf{h}_t =& \mathbf{z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \odot \\ & \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}), \end{aligned} \tag{2}$$

where each $\mathbf{W}$ and each $\mathbf{U}$ are parameters. $\mathbf{h}_t$ is the response of the t^{th} stage. In traditional GRU, the parameters are learned

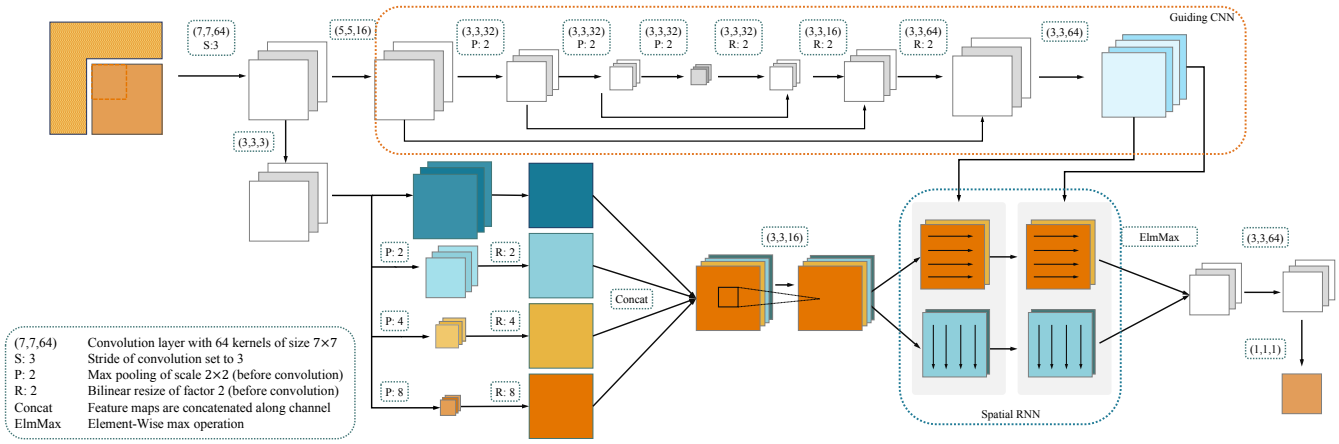

Fig. 4. Architecture of the whole network. Input blocks are handled in a multi-scale scheme using max pooling and bilinear interpolation. The data path on the above is the guiding CNN to extract high-level feature to provide guidance for the training of the spatial RNN on the bottom.

during training and fixed after the training ends. Differently, in our proposed model, we exploit a guiding CNN to generate the parameters. The original formulation in Eq. (1) is now redefined as follows,

$$\tilde{\mathbf{X}}_i^v = \mathcal{F}\left(\mathbf{X}_i^v, \mathbf{X}_{i-1}^v, \phi(\mathbf{X}, \theta')^v\right), \tag{3}$$

where $\mathbf{X}$ is the prediction context. Function ϕ is approximated using a trainable CNN with parameters θ'.

Since the convolutional part does not produce any target pixel for the prediction signals, the problem of asymmetry can be neglected. Besides, as the CNN is used only for feature extraction rather than pixel level mapping, a shallower structure is used to reduce the training difficulty.

A. Integration with the Codec

The proposed network is integrated into HEVC with RDO. For cases where the original intra prediction in HEVC can perform well enough, the RDO process chooses to predict the block with the original HEVC scheme. For some hard cases where HEVC provides poor rate-distortion cost results, the network will be chosen. One additional flag is needed for each PU to save whether to use the deep learning model or the original HEVC. If the deep mode is chosen, no flag for prediction direction is encoded.

We explore the intra prediction for 8×8 blocks. For each coding unit of size 16×16, which includes 4 PUs, the full context is only available for the first PU and fewer reference blocks are available for the rest. To address this problem, two models are trained separately for these two conditions. The top-left PU is predicted by a full-context model which is trained using full context. The rest is predicted by a 3-block model where the blocks on the above, left and left-above are used as training context.

Rather than trained on original frames or images, the model is trained using reconstructed data. This is to prevent a possible distribution difference between the reconstructed data and the original image signals. As when decoding, only reconstructed blocks are available. When the frames are encoded with quantization, an additional noise will be added to the reconstructed signals.

TABLE I
SUMMARIZED COMPARISON WITH THE MODEL IN [7].

Sequence	BD-Rate	
	Ours	Li *et al.* [7]
Class A	-0.8%	-1.4%
Class B	-1.5%	-1.3%
Class C	-1.0%	-0.5%
Class D	-0.9%	-0.5%
Class E	-1.7%	-0.9%
Average	-1.2%	-0.9%

IV. Experimental Results

A. Training Settings

As video frames are much alike in one sequence, they are not ideal training data for intra prediction models, which embraces diversity. The training data is generated from high-resolution images provided in [13]. These images have not been lossily compressed, which makes them artifacts free. To make the model adapt to various resolution, we cropped and downsampled the images to 3 scales, namely 1792×1024, 1344×768, and 896×512. The images are encoded using HEVC with Quantization Parameter (QP) set to $22, 27, 32, 37$ respectively and we use the reconstructed blocks in the decoding process to form the training pairs. We randomly sample about 3,000,000 samples for the training.

Training the model using pairs with high QP settings can enhance the ability for models to overcome the influence of the quantization noise, but these training pairs are less expressive in terms of the original mapping from the reference signals to the predicted signals. Our mixed training set balances the robustness and expressiveness of the spatial RNN model. In this way, the model can be trained with robustness to noise without harming its expressiveness for spatial mapping. We

TABLE II
COMPARISON WITH HEVC, EVALUATED IN BD-RATE.

Class	Sequence	BD-Rate		
		Y	U	V
Class A	Traffic	-1.3%	-1.0%	-0.9%
	PeopleOnStreet	-1.1%	-0.8%	-0.8%
	NebutaFestival	-0.2%	0.0%	0.0%
	SteamLocomotiveTrain	-0.6%	0.0%	0.0%
	Class A Average	-0.8%	-0.5%	-0.4%
Class B	Kimono	-2.8%	-2.2%	-1.7%
	ParkScene	-1.7%	-0.3%	-0.2%
	Cactus	-1.2%	-0.5%	-0.7%
	BasketballDrive	-1.0%	-0.4%	-1.1%
	BQTerrace	-1.0%	0.6%	-0.7%
	Class B Average	-1.5%	-0.6%	-0.9%
Class C	BasketballDrill	-1.0%	-0.2%	0.7%
	BQMall	-0.8%	-0.2%	-0.6%
	PartyScene	-1.0%	-0.4%	-0.7%
	RaceHorses	-1.1%	-0.6%	-1.2%
	Class C Average	-1.0%	-0.4%	-0.5%
Class D	BasketballPass	-0.8%	-0.3%	0.7%
	BQSquare	-0.6%	-1.8%	-0.1%
	BlowingBubbles	-1.0%	-0.7%	0.0%
	RaceHorses	-1.1%	-0.9%	-0.6%
	Class D Average	-0.9%	-0.9%	0.0%
Class E	Johnney	-2.2%	-0.2%	-1.6%
	FourPeople	-1.5%	-0.8%	-0.1%
	KristenAndSara	-1.3%	-0.6%	-1.0%
	Class E Average	-1.7%	-0.5%	-0.9%
Average		-1.2%	-0.6%	-0.5%

use the mean square error (MSE) objective function to jointly train the spatial RNN and the guiding CNN. The network is trained using Stochastic Gradient Descent (SGD) with an initial learning rate 0.01. The learning rate is set to decay exponentially with factor 0.7 for every 10 epochs and the whole training last for 50 epochs.

B. Performance Evaluation

We implement the network into HEVC Test Model (HM) 16.15. The intra main configuration in the common test conditions (CTC) [14] is used. The anchor and proposed method only allow CU size of 16×16 and forced to do a split. That is, each PU is restricted to have the size of 8×8. The Most-Possible-Mode (MPM) is disabled in the test model. The QP is set to $[22, 27, 32, 37]$. When testing on videos, encoders with different QPs share the same models. The rate-distortion performance is measured using BD-Rate [15]. As intra prediction handle each frame separately, a comparison can be made on a relatively small number of frames. For each video sequence, the first 5 frames are tested. The result for each testing sequence is listed in table II. We also compare our result with the work in [7] which utilize an FC network for intra prediction. The result is shown in table I.

It can seem in Table II and Table I that, our proposed model brings better rate-distortion performance than HEVC, especially for Class E. The reason for this is that this scale mostly matches our training set distribution and is partially due to the video sequences in Class E has large smooth areas and our network-based intra predictor can further save bits. For videos with different resolution, our model can also provide a robust and satisfactory result and brings bitrate saving compared with previous methods.

V. CONCLUSION

In this paper, we propose an optimized intra prediction method for video coding. We integrate the spatial RNN and guiding CNN into HEVC and enhance the intra predictor. The RNN architecture solves the problem of asymmetry which may challenge other network structure in intra prediction task. Experimental results show improvement in rate-distortion performance compared with HEVC and previous deep network-based approaches.

REFERENCES

[1] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[2] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[3] J. Li, B. Li, J. Xu, and R. Xiong, "Intra prediction using multiple reference lines for video coding," in *Proc. of Data Compression Conference*, 2017.

[4] ——, "Efficient multiple line-based intra prediction for HEVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 4, pp. 947–957, 2016.

[5] X. Zhang, W. Yang, Y. Hu, and J. Liu, "DMCNN: Dual-domain multi-scale convolutional neural network for compression artifacts removal," in *Proc. of IEEE International Conference on Image Processing*, 2018.

[6] S. Xia, W. Yang, Y. Hu, S. Ma, and J. Liu, "A group variational transformation neural network for fractional interpolation of video coding," in *Proc. of Data Compression Comference*, 2018.

[7] J. Li, B. Li, J. Xu, and R. Xiong, "Intra prediction using fully connected network for video coding," in *Proc. of IEEE International Conference on Image Processing*, 2017.

[8] N. Yan, D. Liu, H. Li, and F. Wu, "A convolutional neural network approach for half-pel interpolation in video coding," in *Proc. of IEEE International Symposium on Circuits and Systems*, 2017.

[9] Z. Liu, X. Yu, S. Chen, and D. Wang, "CNN oriented fast HEVC intra CU mode decision," in *Proc. of IEEE International Symposium on Circuits and Systems*, 2016.

[10] W. Cui, T. Zhang, S. Zhang, F. Jiang, W. Zuo, Z. Wan, and D. Zhao, "Convolutional neural networks based intra prediction for HEVC," in *Proc. of Data Compression Conference*, 2017.

[11] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[12] S. Liu, J. Pan, and M.-H. Yang, "Learning recursive filters for low-level vision via a hybrid neural network," in *Proc. of European Conference on Computer Vision*, 2016.

[13] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang *et al.*, "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017.

[14] F. Bossen, "Common test conditions and software reference configurations," in *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, 5th meeting, Jan. 2011*, 2011.

[15] G. Bjontegarrd, "Calculation of average psnr differences between rd-curves," *VCEG-M33*, 2001.