Combining Intra Block Copy and Neighboring Samples Using Convolutional Neural Network for Image Coding

Zhaobin Zhang†
Dept. Computer Science Electrical Engineering
University of Missouri Kansas City
Kansas City, MO, USA
zzktb@mail.umkc.edu

Yue Li‡
Dept. Electronic Engineering Information Science
University of Science and Technology of China
Hefei, Anhui, China
lytt@mail.ustc.edu.cn

Li Li
Dept. Computer Science Electrical Engineering
University of Missouri Kansas City
Kansas City, MO, USA
lil1@umkc.edu

Zhu Li
Dept. Computer Science Electrical Engineering
University of Missouri Kansas City
Kansas City, MO, USA
lizhu@umkc.edu

Shan Liu
Media Lab
Tencent America
Palo Alto, CA, USA
shanl@tencent.com

***Abstract*—Intra prediction is an essential step to remove spatial redundancy in image coding. How to best predict current block given surrounding pixels is the key efficiency challenge. Inspired by the recent success in applying deep learning to image/video coding systems, we propose an intra prediction method by combining intra block copy and neighboring samples using convolutional neural networks. A novel CNN is developed to further exploit the spatial correlation. Instead of only considering local information, the proposed method can infer the current block via fusing the non-local recurrent features, which is captured by intra block copy, with the local samples located at the left and above boundaries of current block. We also investigate how the performance is affected by the way of fusing IBC and reference boundary pixels. In additional, training data pre-processing is studied to enable the CNN with a better learning capability. Simulation results yield promising coding gain and indicate great potential ability that CNN can be used for next generation video coding framework.**

***Index Terms*—Convolutional neural network (CNN), Intra prediction, High Efficiency Video Coding (HEVC), Image coding, Intra block copy.**

I. Introduction

High Efficiency Video Coding (HEVC) [1] has been the state-of-the-art video coding scheme for the past decade. As one of the key components, intra prediction is the crucial step to remove spatial redundancy in a frame. Significant improvement has been achieved compared with last generation video coding standard H.264/AVC.

HEVC intra prediction follows similar idea from H.264/AVC, which is based on directional extrapolation but enabling more intra prediction angles. There are Planar, DC and 33 angular intra prediction modes in HEVC. Planar and DC are designed for large flatten areas and slow-changing areas respectively. Angular modes target at predicting image blocks possessing dominant directional textures by extrapolating reference pixels along the certain direction. The existing intra prediction performs well in predicting the local, continuous and directional image features with a low computational complexity.

Some advanced intra prediction schemes have also been extensively investigated. Intra block copy is one of efficient tools among them, it generates the intra prediction by performing block matching in the reconstructed area of current frame. Since intra block copy allows searching in the whole reconstructed area, it is especially good at predicting non local, recurrent image patterns. Actually, IBC has already been adopted in HEVC extension for screen content coding [2].

Recently, deep learning has shown superior capability in various tasks, such as image classification [3], image restoration [4]. There are also some recent works applying deep learning on video coding. [5] proposed a down-/up-sampling-based coding scheme using CNN for intra coding. Significant coding gain has been observed. [6] tried to fit a fully connected network on multiple reference pixels and achieved promising results.

Motivated by the recent advances of deep learning in video coding and the efficiency of IBC, we devise a CNN-based Intra Prediction method, termed CIP which tries to learn a model better characterizing the correlation between reconstructed pixels and current block. The proposed CIP is fed with multiple inputs, including both the neighboring reference pixels and the best prediction found by IBC. We also investigated different fusion methods to better leverage

†Part of this work was done while Zhaobin Zhang was on his internship in Tencent America.

‡This work was done when Yue li was an visiting scholar in University of Missouri Kansas City.

extracted features. Data preprocessing is very crucial to train a clean model, therefore a threshold is set to remove distractors from raw IBC blocks. The experimental results show that the proposed scheme achieves an average of 1.3% bitrate saving on luma component for HEVC test sequences under all intra configuration compared with HEVC 16.0 anchor.

The rest of this paper is organized as follows. The proposed method is elaborated in Section II. Experimental results and analysis are described in Section III. Section IV concludes the paper.

II. The Proposed Method

With support of modern advanced mobile devices, multimedia been dramatically increasing over the Internet and video is undoubtedly the dominant Internet traffic through the world. The rich contextual information existing in natural video poses unique challenges on streaming such huge data. In existing HEVC implementations, a large portion of spatial correlation cannot be captured due to the limits of simple interpolation solution. Intra block copy has been demonstrated the effectiveness in estimating recurrent patterns, e.g., screen content coding. In addition, CNN has shown superior power in various tasks. It is natural to try combining them together by CNN. As CNN is adept in exploiting complicated hidden features, we expect it performs a better job in intra prediction. First, the CIP architecture is introduced. Second, the training procedure and hyper-parameter tuning are detailed. Finally, how the proposed method is integrated in HEVC reference software is elaborated.

A. CIP Architecture

With support of modern GPUs, current CNN has been developed to contain more layers and complicated architectures. However, it is not applicable for video coding which requires higher time efficiency. To balance the trade-off between performance and time complexity, we start with a relatively neat design. The proposed CIP architecture is illustrated in Fig. 1. The framework consists of two parts, fully-connected (FC) layers for reference boundary pixels feature extraction and convolutional neural networks for fusion pixels reconstruction.

For current ground truth block Y_0 with size $N \times N$, the proposed CIP aims at learning a projection from $\{R, P\}$ to Y_0, where R is the $2N+1$ boundary reference pixels and P is the intra block copy prediction. Intra block copy is obtained through searching in a wider range from the reconstructed regions while the boundary reference pixels can be leveraged to extract local prediction. Intuitively, convolutional neural networks are expected to learn a better projection by combining the local and non-local information. To make the network easier to train, the pixel values are normalized to range [0, 1]. Let us denote the depth of the fully-connected layers and that of the convolutional neural networks as d_f and d_c respectively. The input of the FC layers is the $2N+1$ boundary reference pixels and outputs a vector contains N^2 pixel values which will be reshaped to a $N \times N$ block. For the ith FC layer, its output is a vector in K_i-dimensional, and it is calculated as.

$$F_i^F(x) = W_i^F \cdot x_i^F + b_i^F, \quad 1 \le i \le d_f \tag{1}$$

where W_i^F and b_i^F represents weights and biases. x_i^F is the input vector of the ith FC layer. When i equals to one, x_1^F is the input reference pixels in $2N+1$ dimensional. Each FC layer is followed by a non-linear activation layer.

The output of FC layers is reshaped to a $N \times N$ block M. Element-wise summation is applied on M and intra block copy P followed by convolutional neural networks. Each convolutional layer is followed by a non-linear activation layer which is not counted. Given the previous layer's output x_j^C, current layer feature map F_j^C is calculated as follows.

$$F_j^C = W_j^C \cdot x_j^C + b_j^C, \quad 1 \le j \le d_c \tag{2}$$

where W_j^C and b_j^C are the weights and biases of current layer. x_j^C is the output of previous layer. When $j = 1$, namely the first layer, the input should be the summation of FC output M and intra block copy P. Both FC layers and convolutional layers adopt Rectified linear unit (ReLU) as the activation function. Residual learning is adopted in the proposed method due to its fast convergence property. Intra block copy is added to the end of CNN.

To better control the trade-off between the IBC prediction accuracy and the computational complexity. The process of IBC is parametrized by search range r. Δx and Δy denote the displacements from current block coordinates along x and y axis respectively. The values of Δx and Δy should guarantee the IBC reference block is within current available pixel regions.

$$r \le |\Delta x| + |\Delta y| \tag{3}$$

B. Training

The training process is actually estimating the network parameter $\Theta = \{W^F, b^F, W^C, b^C\}$ with predefined loss function. The objective of this network is to learn the end-to-end mapping from (R, P) to Y_0, i.e.,

$$f(R, P, \Theta) = Y \tag{4}$$

This is achieved by minimizing the loss between the predicted block Y and the corresponding ground truth block Y_0. Euclidean loss is adopted in this work due to its simplicity and popularity. To avoid over-fitting during the training procedure, regularization term is added to the loss function. Suppose there are totally S training pairs, the loss function is formulated as follows.

$$L(\Theta) = \frac{1}{2S} \sum_{s=1}^{S} \|Y^s - Y_0^s\|_2^2 + \frac{\lambda}{2} \|\Theta\|_2^2 \tag{5}$$

where λ is the regularization term and is set to 10^{-4} in implementation. Stochastic gradient descent (SGD) is utilized to minimize the loss while training. SGD updates the parameter set Θ by combining current gradient and previous parameter

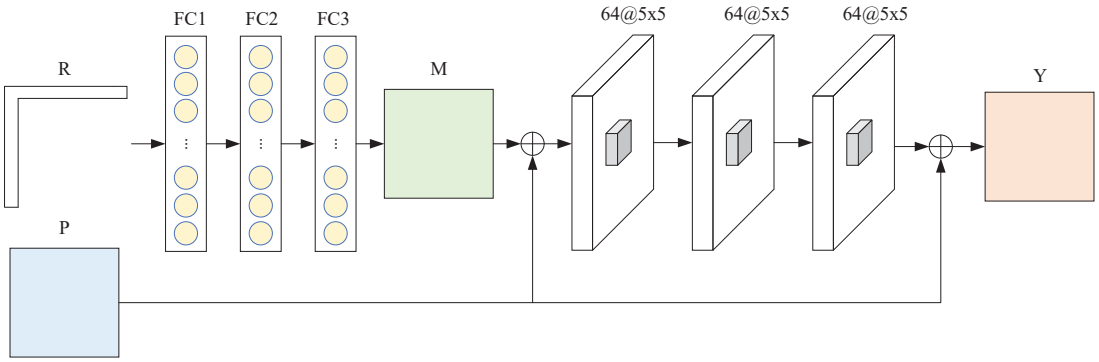

Fig. 1. The proposed CIP accepts both boundary reference pixels R and intra block copy P as inputs. Fully-connected layers are performed on the boundary reference pixels and the output is reshaped to a block M followed by element-wise summation with the intra block copy. Then it is fed into convolutional neural networks which try to fit an end-to-end model using residual learning.

update. Specifically, update at iteration $t+1$ can be expressed as:

$$V_{t+1} = uV_t - \alpha \nabla L(\Theta) \tag{6}$$

$$\Theta_{t+1} = \Theta_t + V_{t+1} \tag{7}$$

where $\nabla L(\Theta)$ are gradients with respect to the parameters Θ to be updated. V_t is previous parameter update. α is the learning rate and u is the momentum. Θ is initialized by random Gaussian distribution with zero mean and standard deviation of 1. Momentum u is set as 0.9 and the learning rate α is set to decay exponentially from 10^{-4} to 10^{-9} by a factor of 10^{-1}.

C. Integration in HM

In this work, we design the CIP model cooperate with existing directional intra prediction model in HEVC coding framework, and rate-distortion optimization is used to choose the optimal model. A binary flag will be transmitted to indicate whether CIP is adopted. As shown in Fig. 2, the proposed method is plugged after the existing intra prediction process. An optimal intra prediction mode will be obtained after the 35 intra prediction mode decision. The proposed CIP is performed after that and the optimal prediction mode is updated accordingly. If CIP mode is adopted, corresponding motion vector will be transmitted to decoder. The decoder will first check whether CIP is used. If CIP is used, motion vector will be decoded followed by the CIP model, otherwise conventional directional intra decoding process is performed.

III. Experiments

The proposed method is integrated into HM 16.0 [7] and the results are compared with HM-16.0 anchor. The preliminary experiments will only show block size of 16×16. In this section, experimental settings are introduced first. Training data derivation is detailed afterwards. Finally, the experimental results and analysis are shown.

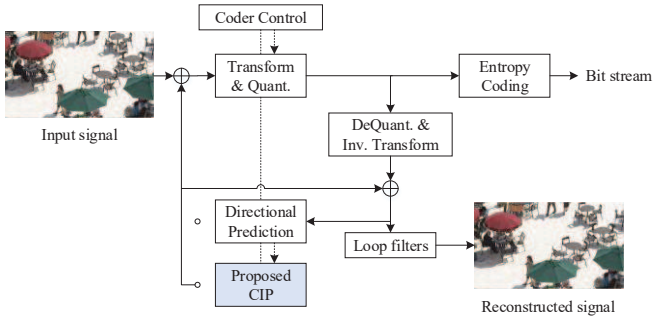

Fig. 2. Integration of the proposed CIP scheme in HM software

A. Experimental Settings

In HEVC, the block-based intra prediction follows a quadtree pattern resulting the CU size ranging from 8 to 64. As block size 64 is too large to find an ideal intra block copy prediction, we constraint the block size ranging from 8 to 32. The preliminary work only implements block size of 16 and the other sizes will be left for future work. All-intra configuration suggested by HEVC common test condition is adopted. Quantization parameter (QP) is set to {22, 27, 32, 37}. The CIP model is implemented in caffe [8]. The base learning rate is set to 0.0001 and decay exponentially every 100k iterations. The total number of iterations is 1.5M and it takes about 20 hours on a GTX 1080Ti GPU.

B. Training Data Derivation

DIV2K 2K resolution high quality image dataset [9] is used to generate the training data. There are 1000 high definition high resolution images among which 800 images for training, 100 images for validation and another 100 images for test. The images content covers a wide range of objects.

Data preparation is to find the input pair $\{R, P\}$. The reference pixels R is extracted from the original image. The IBC search range is empirically set to $r = 128$ to balance the time complexity and matching accuracy. The IBC block is searched pixel by pixel within the available pixels according to Eq. 3. The difference is measured by Sum of Absolute

Difference (SAD) and the block which has the minimum SAD from the ground truth is selected as current IBC prediction. To refine the IBC training samples, mean square error (MSE) is used to measure the distance between IBC to corresponding ground truth. We only keep those IBC with a MSE smaller than 0.5, in such a way, about 55% training samples are preserved.

C. Results and Analysis

Three different kinds of fusion methods have been tried in the proposed method, e.g., element-wise addition, 1x1 Convolution and Concatenation. The simulation results indicate that the element-wise addition is the optimal fusion in our method. Therefore, the final results are implemented with element-wise addition.

The BD-Rate reduction of all the test sequences are listed in Table I. In addition, we also show the average results for each class. An average of -1.3%, -1.4% and -1.6% BD-Rate saving are achieved by the proposed method for luma, Cb and Cr component respectively. The peak performance on luma component is on BQTerrace with -3.6% BD-Rate reduction. Overall, the performance of proposed method on chroma component is consistent with that on luma component. And the best performance on Cb and Cr is on sequence BasketballDrive and BQTerrace with -5.7% and -4.5% BD-Rate reduction respectively. The encoding time is much higher than anchor and decoding time is 50% than anchor. Fortunately, the encoding time does not count that much compared with decoding time complexity in video coding. It is noticeable that the proposed method does not perform very well on Class D which might be caused by the training dataset not covering similar content. It should be noticed that this version is our preliminary benchmark and the same CIP model is shared among different block sizes. In addition, different dataset also has different impacts on the final performance. All possible improvement work will be left for the future work.

TABLE I
THE BD-RATE RESULTS OF THE PROPOSED METHOD.

Sequence		BD-Rate		
		Y	U	V
Class A	Traffic	-1.5%	-0.8%	-0.7%
	PeopleOnStreet	-2.1%	-2.8%	-2.6%
	Nebuta	-0.4%	-0.7%	-0.4%
	SteamLocomotive	-0.1%	0.9%	-0.1%
Class B	Kimono	0.3%	-0.2%	-0.3%
	ParkScene	-0.6%	-1.1%	-1.0%
	Cactus	-2.0%	-2.7%	-1.8%
	BQTerrace	-3.6%	-2.5%	-4.5%
	BasketballDrive	-2.8%	-5.7%	-4.2%
Class C	BasketballDrill	-2.8%	-4.5%	-3.6%
	BQMall	-0.8%	0.1%	-0.7%
	PartyScene	-0.7%	0.2%	0.1%
	RaceHorsesC	-0.2%	-0.4%	0.2%
Class D	BasketballPass	-0.7%	-0.7%	-3.4%
	BQSquare	-0.9%	-0.3%	0.1%
	BlowingBubbles	0.2%	-1.0%	-1.0%
	RaceHorses	0.3%	1.3%	0.9%
Class E	FourPeople	-1.3%	0.5%	-0.3%
	Johnny	-2.7%	-2.3%	-3.9%
	KristenAndSara	-2.1%	-4.9%	-2.5%
Class A		-1.1%	-0.8%	-1.0%
Class B		-1.8%	-2.3%	-2.4%
Class C		-1.2%	-1.2%	-1.0%
Class D		-0.2%	-0.2%	-0.8%
Class E		-2.0%	-2.2%	-2.6%
Average		-1.3%	-1.4%	-1.6%
Enc Time		794%		
Dec Time		150%		

IV. CONCLUSION

This paper proposes a novel CNN-based intra prediction scheme which utilizes CNN's superior capability on various tasks to further exploit the spatial correlation in a frame. The proposed CNN is fed with both the neighboring pixels and the best prediction found by intra block copy. Different methods of combining intra block copy and boundary reference pixels have been investigated and element-wise addition is adopted. To make the proposed CNN more tractable and efficient, training data is refined by empirically setting a threshold to remove distractors. The experimental results demonstrate the effectiveness of the proposed scheme and indicate the latent capability of CNN on improving existing coding efficiency. These observations are valuable for the development of next generation video coding tools and yield a lot of interesting research directions.

REFERENCES

[1] G. J. Sullivan, J. R. 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, Dec 2012.

[2] Y. Li, L. Li, D. Liu, H. Li, and F. Wu, "Combining directional intra prediction and intra block copy with block partition for hevc," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 524–528.

[3] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2017.

[4] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," *arXiv preprint arXiv:1801.07892*, 2018.

[5] Y. Li, D. Liu, H. Li, L. Li, F. Wu, H. Zhang, and H. Yang, "Convolutional neural network-based block up-sampling for intra frame coding," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2017.

[6] J. Li, B. Li, J. Xu, R. Xiong, and W. Gao, "Fully connected network-based intra prediction for image coding," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3236–3247, July 2018.

[7] HEVC. (2018) Hevc reference software 16.0. [Online]. Available: https://hevc.hhi.fraunhofer.de/trac/hevc/browser/tags/HM-16.0

[8] B. A. Research. (2018) Caffe. [Online]. Available: http://caffe.berkeleyvision.org/

[9] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 1122–1131.