A HYBRID NEURAL NETWORK FOR CHROMA INTRA PREDICTION

Yue Li[1], *Li Li*[2], *Zhu Li*[2], *Jianchao Yang*[3], *Ning Xu*[3], *Dong Liu*[1], *Houqiang Li*[1]

[1]CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System,
University of Science and Technology of China, Hefei 230027, China
lytt@mail.ustc.edu.cn, {dongeliu,lihq}@ustc.edu.cn
[2]University of Missouri at Kansas City, MO 64111, USA, {lil1,lizhu}@umkc.edu
[3]Snapchat Inc., CA 90291, USA, {jianchao.yang,ning.xu}@snapchat.com

ABSTRACT

For chroma intra prediction, previous methods exemplified by the Linear Model method (LM) usually assume a linear correlation between the luma and chroma components in a coding block. This assumption is inaccurate for complex image content or large blocks, and restricts the prediction accuracy. In this paper, we propose a chroma intra prediction method by exploiting both spatial and cross-channel correlations using a hybrid neural network. Specifically, we utilize a convolutional neural network to extract features from the reconstructed luma samples of the current block, as well as utilize a fully connected network to extract features from the neighboring reconstructed luma and chroma samples. The extracted twofold features are then fused to predict the chroma samples–Cb and Cr simultaneously. The proposed chroma intra prediction method is integrated into HEVC. Preliminary results show that, compared with HEVC plus LM, the proposed method achieves on average 0.2%, 3.1% and 2.0% BD-rate reduction on Y, Cb and Cr components, respectively, under All-Intra configuration.

Index Terms— Chroma intra prediction, convolutional neural network, fully connected network, hybrid neural network.

1. INTRODUCTION

Recently, the constantly increasing video resolution raises more and more severe challenge to video compression techniques. High Efficiency Video Coding (HEVC), which is finalized by experts from the Joint Collaborative Team on Video Coding, is the state-of-the-art video coding standard [1]. Compared to its predecessor, H.264 [2], HEVC achieves approximately the same subjective quality with around 50% less bit rate. Despite the superior performance of HEVC, efficiency of video compression techniques still cannot meet the bandwidth and storage demand for the explosively growing video contents. Advanced video compression methods are in urgent need.

The chroma intra prediction in HEVC consists of three directional prediction modes, one direct current (DC) mode and one planar mode [3]. Directional prediction modes generate prediction of the current block by a directional extrapolation from the neighboring reconstructed samples at the upper and left boundaries. During the development of HEVC, some advanced chroma intra prediction methods have been investigated. For example, considering the built-in coding order of the three channels, i.e. luma component is compressed and reconstructed before chroma components, Kim *et al.* proposed the Linear Model method (LM) for chroma intra prediction [4]. LM assumes a linear correlation between the luma and chroma components in a coding block, and predicts chroma samples from reconstructed luma samples using linear regression. The regression parameters are not transmitted, but rather derived from the reconstructed neighboring luma and chroma samples. LM reported promising results, which inspired following-up researches on improving LM [5, 6, 7]. However, these subsequent improved versions all inherit the linear correlation assumption, which is inaccurate for complex image content or large coding blocks.

Recent years have witnessed a great success of convolutional neural network (CNN) in computer vision and image processing, such as image recognition [8], image super-resolution [9] and image recoloring [10], etc. For example, in image recoloring, the objective is to infer color information from grayscale images. Iizuka *et al.* proposed an end-to-end CNN for automatic recoloring of grayscale images [10]. They used a combination of global image priors, which are extracted from the entire image, and local image features, which are computed from small image patches, in the proposed CNN.

Inspired by the image recoloring using CNN, we would like to investigate a neural network-based approach for

This work was supported by the National Program on Key Basic Research Projects (973 Program) under Grant 2015CB351803, the Natural Science Foundation of China (NSFC) under Grants 61772483, 61390514, and 61331017, a gift grant by Snapchat Inc., and CSC scholarship. *(Corresponding author: Zhu Li and Dong Liu.)*

chroma intra prediction. However, different from image recoloring where luma information of the entire image is available, chroma intra prediction in video coding usually operates at block level. Thus, the CNN needs redesign. We are then motivated by the LM method that exploits not only cross-channel correlation but also spatial correlation, in accordance to the block-based coding scheme.

In this paper, we propose a chroma intra prediction method using a hybrid neural network. Specifically, we utilize a convolutional neural network to extract features from the reconstructed luma samples of the current block, as well as utilize a fully connected network to extract features from the neighboring reconstructed luma and chroma samples. The extracted twofold features are then fused to predict the chroma samples. The fusion is empirically designed as element-wise product in network. The two chroma components–Cb and Cr–are predicted in a single network to reduce complexity. To deal with variable block size in HEVC, we adopt the same network structure but empirically set different hyperparameters for different sized blocks. Owing to the powerful learning capacity of neural network, the proposed method outperforms LM and its variants and achieves significant BD-rate reduction.

The remainder of this paper is organized as follows. Section 2 elaborates the structure of the proposed hybrid neural network. Section 3 discusses the integration of the proposed method into HEVC. Experimental results and analyses are presented in Section 4, followed by conclusions in Section 5.

2. THE PROPOSED HYBRID NEURAL NETWORK

Although both LM and the proposed method predict chroma components from luma samples, our prediction pipeline differs from LM in two folds. On the one hand, we adopt fully connected layers to extract useful information from neighboring samples, while LM simply "reuses" the linear regression parameters. On the other, instead of modeling a linear relation between luma and chroma components like in LM, we embrace more complex convolutional layers to model the underlying correlation between the luma and chroma components.

The structure of the proposed hybrid neural network is illustrated in Fig. 1. We will discuss the detailed structure of the fully connected layers, the convolutional layers and the fusion layer in the following.

2.1. Convolutional layers

We apply two convolutional layers to extract feature maps from the reconstructed luma samples of the current block. In this section, we take the block of 32×32 luma samples and YUV 4:2:0 format as example. The cases of other block sizes are similar and will be detailed in Section 3. The reconstructed luma samples are down-sampled by a factor of 2 to have the same resolution with the chroma samples, and then fed into the convolutional neural network. Denote the down-sampled luma samples as $\boldsymbol{X} \in R^{16\times16}$, the output of the first convolutional layer (equipped with ReLU non-linearity) will be

$$\boldsymbol{C}_1(\boldsymbol{X}) = \max(0, \boldsymbol{W}_1 * \boldsymbol{X} + \boldsymbol{B}_1) \tag{1}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{B}_1$ represent the convolutional filters and biases of the first layer, respectively, $\boldsymbol{C}_1 \in R^{128\times16\times16}$ indicates the 128 feature maps of the first layer, and $*$ stands for convolution. Note that proper padding is applied to ensure the feature maps are of the same resolution as input.

The feature maps extracted by the first layer are taken as input to the second layer, while the second layer is represented as

$$\boldsymbol{C}_2(\boldsymbol{X}) = \begin{cases} \boldsymbol{C}_{21}(\boldsymbol{X}) = \max(0, \boldsymbol{W}_{21} * \boldsymbol{C}_1(\boldsymbol{X}) + \boldsymbol{B}_{21}) \\ \boldsymbol{C}_{22}(\boldsymbol{X}) = \max(0, \boldsymbol{W}_{22} * \boldsymbol{C}_1(\boldsymbol{X}) + \boldsymbol{B}_{22}) \end{cases} \tag{2}$$

where $(\boldsymbol{C}_{21}(\boldsymbol{X}), \boldsymbol{C}_{22}(\boldsymbol{X}))$, $(\boldsymbol{W}_{21}, \boldsymbol{W}_{22})$, $(\boldsymbol{B}_{21}, \boldsymbol{B}_{22})$ are the extracted feature maps, convolutional filters, and biases, respectively. Here we explicitly employ two sets of convolutional kernels with different kernel sizes in the second layer, since the combination of different sized kernels is capable in effectively aggregating multi-scale information [11, 12], The output feature maps $\boldsymbol{C}_{21}$ and $\boldsymbol{C}_{22}$ by different sized kernels are directly concatenated into $\boldsymbol{C}_2 \in R^{128\times16\times16}$, which is then fed into the next layer.

The third convolutional layer has a similar mapping function to the second convolutional layer except that the multi-scale kernel sizes are different. Finally, the predicted chroma samples are derived using the fourth convolutional layer. Note that the two chroma components are predicted using a single network to reduce complexity.

2.2. Fully connected layers

As depicted in Fig. 1, three successive fully connected layers are adopted to extract information from the neighboring reconstructed samples. The neighboring reconstructed luma samples are down-sampled by a factor of 2, and the samples at the upper and left boundaries, in total 33 samples are used as input. Similarly, the neighboring reconstructed chroma samples at the upper and left boundaries, in total 33×2 samples, are also used as input. Thus the input consists of 99 samples, and is denoted by $\boldsymbol{Y} \in R^{99}$. The output of the last fully connected layer is a 128-dimensional feature vector, denoted as $\boldsymbol{F}_3 \in R^{128}$.

2.3. Fusion layer

The fusion layer integrates the neighboring information, i.e. $\boldsymbol{F}_3$, with the feature maps of the luma samples of the current block, i.e. $\boldsymbol{C}_2$. First, we tile the vector $\boldsymbol{F}_3$ into the matrix

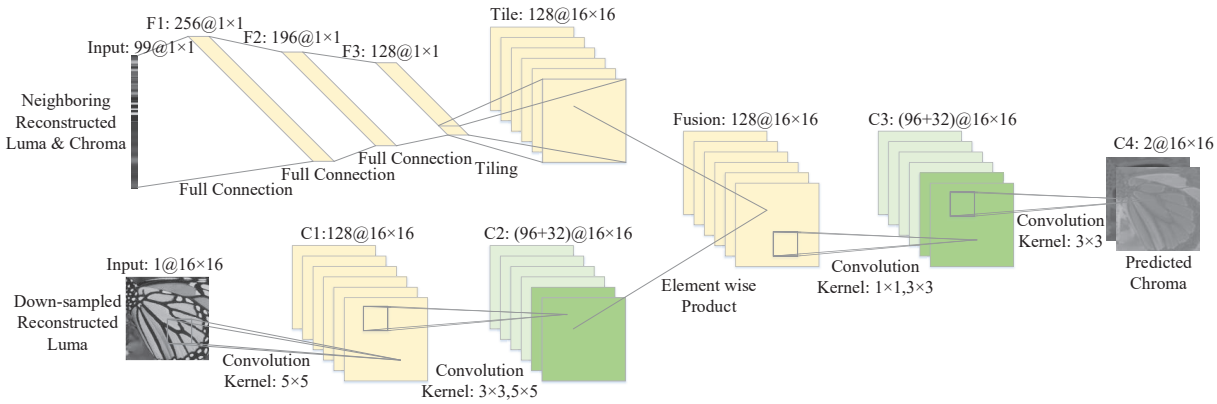

Fig. 1. The structure of the proposed hybrid neural network, where the numbers and sizes (lengths) of feature maps (vectors) are shown on top of them, and "+" stands for concatenation of feature maps.

$\boldsymbol{T} \in R^{128\times16\times16}$:

$$\boldsymbol{T}^i_{u,v} = \boldsymbol{F}^i_3, i \in [1, 128], u, v \in [1, 16] \quad (3)$$

Then we perform the fusion by element-wise product. According to our empirical results, this simple operation is easy to train and works well:

$$\boldsymbol{Fusion}^i_{u,v} = \boldsymbol{T}^i_{u,v} \times (\boldsymbol{C}_2)^i_{u,v}, i \in [1, 128], u, v \in [1, 16] \quad (4)$$

Another interpretation of the fusion operation is a piecewise linear modeling of the correlation between the luma and chroma components in the embedded feature space.

3. INTEGRATION INTO HEVC

3.1. Dealing with variable block size

In HEVC, intra prediction operates according to the transform block (TB) size [1]. Since the TB sizes vary from 4×4 up to 32×32, and the chroma TB size is half of the luma TB size (YUV 4:2:0), the prediction sizes of chroma components actually include 4×4, 8×8 and 16×16. To handle input blocks with different sizes, we use the same network structure as shown in Fig. 1 but with different hyperparameters for different sized blocks.

While there are several kinds of hyperparameters, such as the amount of feature maps, the size of kernels, etc, to consider when dealing with different sized blocks, our preliminary results suggest that the amount of feature maps (or nodes in the fully connected layers) is the most important factor. Therefore, we fix the other hyperparameters, and attempt to adjust the proper amount of feature maps (nodes) for each network. The final settings are shown in Table 1. We empirically set fewer feature maps (nodes) for smaller sized blocks, since small blocks typically contain relatively simple content for which simple network with less parameters can capture the underlying correlation between the luma and chroma components.

Table 1. The amount of feature maps (nodes) in each layer of the networks for different sized blocks: 4×4, 8×8 and 16×16.

Type	# Feature Maps (Nodes)		
	4×4	8×8	16×16
conv.	32	64	128
conv.	24,8	48,16	96,32
conv.	24,8	48,16	96,32
conv.	2	2	2
FC	64	128	256
FC	49	98	196
FC	32	64	128

3.2. Integration into HEVC

To evaluate the proposed chroma intra prediction method, two new prediction modes including the LM and the proposed one are integrated into HEVC for chroma component intra coding, in addition to the existing ones. When the proposed mode is used, the Cb and Cr components are predicted simultaneously using the trained hybrid neural network. The encoder will choose the best prediction mode for each prediction unit, according to the minimum rate-distortion cost criterion.

4. EXPERIMENTAL RESULTS

In order to train the proposed hybrid neural network, we use the DIV2K [13], which is a newly released high-quality high-resolution image dataset containing 800 training images, to generate training data. The Caffe software [14] is used to train CNN models. We do not distinguish QP when generating training data, so in total we have 3 models corresponding to the 3 possible TB sizes.

We have integrated both LM and the proposed method

Table 2. BD-rate results of the proposed method. We compare HM plus LM and the proposed method, with the scheme of HM plus LM.

Class	Sequence	Y	U	V
Class A	Traffic	–0.0%	–2.1%	–0.7%
	PeopleOnStreet	–0.2%	–2.4%	–2.5%
	Nebuta	–0.6%	–9.2%	–0.8%
	SteamLocomotive	–0.0%	–9.1%	1.5%
Class B	Kimono	–1.2%	–5.4%	0.1%
	ParkScene	–0.7%	–8.9%	–0.7%
	Cactus	–0.0%	–4.0%	–3.8%
	BQTerrace	–0.2%	0.4%	–1.9%
	BasketballDrive	–0.2%	–3.6%	–4.2%
Class C	BasketballDrill	–0.3%	–1.6%	0.7%
	BQMall	–0.1%	–5.7%	–3.4%
	PartyScene	–0.2%	–4.6%	–0.9%
	RaceHorsesC	–0.1%	0.6%	–0.3%
Class D	BasketballPass	–0.9%	–1.3%	–4.2%
	BQSquare	0.4%	0.6%	–0.3%
	BlowingBubbles	0.4%	–3.2%	–7.8%
	RaceHorses	0.4%	–2.0%	–1.8%
Class E	FourPeople	–0.1%	–1.8%	–0.2%
	Johnny	0.6%	0.4%	–3.8%
	KristenAndSara	–0.4%	1.2%	–5.2%
Average		–0.2%	–3.1%	–2.0%

Table 3. Comparison with the results in [7], both results are anchored on the scheme of HM plus LM.

Class	The results in [7]			Our results		
	Y	U	V	Y	U	V
Class A	0.0%	–2.2%	–0.5%	–0.2%	–5.7%	–0.6%
Class B	–0.2%	–1.2%	–0.9%	–0.5%	–4.3%	–2.1%
Class C	–0.1%	–1.2%	–1.3%	–0.2%	–2.8%	–1.0%
Class D	–0.2%	–0.8%	–1.9%	0.1%	–1.5%	–3.5%
Class E	–0.2%	–0.9%	–1.6%	0.1%	–0.1%	–3.1%
Average	–0.2%	–1.2%	–1.3%	–0.2%	–3.1%	–2.0%

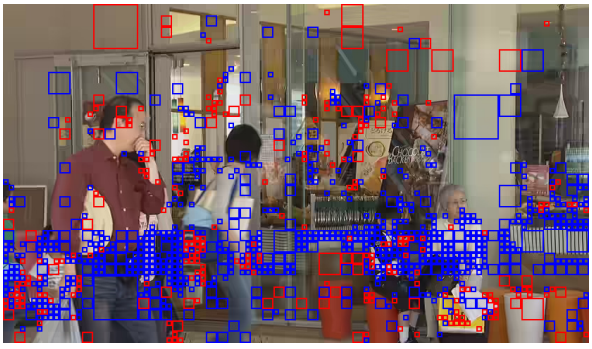

Fig. 2. This figure shows the CUs that choose different modes. CUs with blue boundary are predicted using the proposed method, CUs with red boundary are predicted with LM.

into the HEVC reference software–HM version 12.0[1]. For experiments, we use the all-intra configuration suggested by the common test conditions [15]. The QP is set to 27, 32, 37, 42. We adopt BD-rate [16] to evaluate the relative compression efficiency. Test sequences include 20 video sequences of different resolutions known as Classes A, B, C, D, E [15]. Class F consists of screen content videos and thus is excluded as the CNN models are trained with natural images.

The overall results are summarized in Table 2. Here, we compare the scheme of HM plus LM and the proposed method, with the scheme of HM plus LM. That is, the gain is achieved by our method on top of LM. It can be observed that our proposed method achieves on average 0.2%, 3.1%, and 2.0% BD-rate reduction on Y, U, and V, respectively. Reasonably the BD-rate is more significant for the chroma components. It is worth noting that the proposed method performs especially well on Classes A and B, which we conjecture is due to the similar resolutions of the Classes A, B and the training images.

To the best of our knowledge, the method in [7] achieves the best performance among those improved versions of LM. In Table 3, we compare our results with those reported in [7] under the same experimental setting. It can be observed that the proposed method outperforms the method in [7] with a large margin, especially for high-resolution sequences, i.e. Classes A and B.

For visual inspection, Fig. 2 displays the selected chroma prediction modes for the sequence `BQMall`. We can observe that our proposed mode is selected mostly for the regions with rich textures or structures. Also, our proposed mode can be selected for quite large blocks but LM is mostly used for smaller blocks. Such results demonstrate that our proposed method can better predict chroma components where the simple linear correlation assumption does not hold true.

5. CONCLUSION

We have proposed a hybrid neural network, which consists of a fully connected network and a convolutional neural network, for chroma intra prediction. The feature vector computed by the fully connected network serves as spatial information to weigh the feature maps extracted by the convolutional neural network. Experimental results show the effectiveness of the proposed method. Visual inspection further demonstrates that the proposed method is more efficient than the previous linear correlation assumption-based methods. Currently, our proposed method incurs a high computational complexity due to the neural network. In the future, we will investigate more simple network structures to reduce the complexity while maintain the prediction accuracy.

[1] `https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-12.0/`

6. REFERENCES

[1] Gary J Sullivan, Jens Ohm, Woo-Jin Han, and Thomas 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] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay 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.

[3] Jani Lainema, Frank Bossen, Woo-Jin Han, Junghye Min, and Kemal Ugur, "Intra coding of the HEVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1792–1801, 2012.

[4] Jungsun Kim, Seung-Wook Park, Younghee Choi, Yong-Joon Jeon, and Byeong-Moon Jeon, "New intra chroma prediction using inter-channel correlation," JCTVC-B021, presented at the 2nd meeting of Joint Collaborative Team on Video Coding (JCT-VC), Geneva, Switzerland, July 2010.

[5] Christophe Gisquet and Edouard François, "Model correction for cross-channel chroma prediction," in *DCC*, 2013, pp. 23–32.

[6] Xingyu Zhang, Christophe Gisquet, Edouard Francois, Feng Zou, and Oscar C Au, "Chroma intra prediction based on inter-channel correlation for HEVC," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 274–286, 2014.

[7] Tao Zhang, Xiaopeng Fan, Debin Zhao, and Wen Gao, "Improving chroma intra prediction for HEVC," in *ICMEW*, 2016, pp. 1–6.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, 2014, pp. 184–199.

[10] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa, "Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 110, 2016.

[11] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411–426, 2007.

[12] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[13] Radu Timofte, Eirikur Agustsson, Luc Van Gool, et al., "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *CVPRW*, 2017, pp. 1110–1121.

[14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimedia*, 2014, pp. 675–678.

[15] Frank Bossen, "Common test conditions and software reference configurations," JCTVC-H1100, presented at the 8th meeting of Joint Collaborative Team on Video Coding (JCT-VC), San Jose, CA, USA, February 2012.

[16] Gisle Bjontegaard, "Calcuation of average PSNR differences between RD-curves," VCEG-M33, presented at the 13th meeting of ITU-T Video Coding Experts Group (VCEG), Austin, Texas, USA, April 2001.