This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology

Effective Data Driven Coding Unit Size Decision Approaches for HEVC INTRA Coding

Yun Zhang, *IEEE Senior Member*, Zhaoqing Pan, *IEEE Member*, Na Li, Xu Wang, *IEEE Member*
Gangyi Jiang, *IEEE Member*, and Sam Kwong, *IEEE Fellow*

***Abstract*—High Efficiency Video Coding (HEVC) INTRA coding improves compression efficiency by adopting advanced coding technologies, such as multi-level quad-tree block partitioning and up to 35-mode INTRA prediction. However, it significantly increases the coding complexity, memory access and power consumption, which goes against its widely applications, especially for ultra-high definition and/or mobile video applications. To tackle this problem, we propose an effective data driven Coding Unit (CU) size decision approaches for HEVC INTRA coding, which consists of two stages of Support Vector Machine based fast INTRA CU size decision schemes at four CU decision layers. At the first stage classification, a three output classifier with offline learning is developed to early terminate the CU size decision or early skip checking the current CU depth. As for the samples that neither early skipped nor early terminated, the second stage of binary classification, which learns online from previous coded frames, is proposed to further refine the CU size decision. Representative features for the CU size decision are explored at different decision layers and stages of classifications. Finally, the optimal parameters derived from the training data are achieved to reasonably allocate complexity among different CU layers at given total rate-distortion degradation constraint. Extensive experiments show that the proposed overall algorithm can achieve 27.95% to 80.53% and 52.48% on average complexity reduction for the CU size decision as compared with the original HM16.7 model. Meanwhile, the average Bjonteggard delta peak-signal-to-noise ratio degradation is only -0.08 dB, which is negligible. The overall performance of the proposed algorithm outperforms the state-of-the-art benchmark schemes.**

***Index Terms*—High Efficiency Video Coding, INTRA coding, CU Size Decision, Machine Learning.**

I. Introduction

HIGH Efficiency Video Coding (HEVC) [1] is the latest ongoing video coding standard developed by the Joint Collaborative Team on Video Coding (JCT-VC), which doubles the compression ratio while maintaining the same visual quality as compared with the H.264/MPEG-4 AVC high profile. It demonstrates a great potential for High Definition (HD) video markets, such as IMAX movies, virtual reality, immersive video conferences, surveillance, ultra HDTV and 3D-TV. In order to promote the efficiency of HEVC and its extensions, more modes are developed for different coding modules [1], including flexible block partitioning, INTRA prediction, INTER prediction, transform, filtering parameter decision and Motion Estimation (ME), *etc.*. The above mentioned coding modules try all the mode candidates and select the optimal one with the minimum Rate-Distortion (RD) cost via RD comparison. This “try all and select the best” brutal-force strategy is extremely complex and power/time consuming, which hinders the applications of mobile video, ultra-HD video, video communication and live broadcasting.

Many researchers have been involved in developing fast video coding algorithms. The RD cost distribution [2] and spatial-temporal correlation [3]-[4] were mainly exploited for fast mode decisions in H.264/AVC based codecs. There are seven INTER and three INTRA block modes in H.264/AVC. However, there are three recursive loops in the block partition/modes in HEVC, *i.e.*, Coding Unit (CU), Prediction Unit (PU) and Transform Unit (TU) sizes and dozens of modes in each loop, which significantly increases the total number of mode candidates and their combinations, thus makes the fast mode decision more complicated. Therefore, in addition to spatial, temporal correlation [5] and statistical RD cost based approaches, advanced learning algorithms [6]-[12] were investigated for the mode decision through modeling it as a classification problem. In [6], Zhang *et al.* modelled the HEVC INTER CU size decision as a three-level of hierarchical binary decision problem. Then, joint Support Vector Machine (SVM) based classifier with the optimal weighted factors was proposed to control the risk of false prediction, which reasonably allocates the complexity among different CU depths and achieves significantly coding complexity reduction. Then, in [7]-[8], feature selection algorithms were proposed to select representative features for the SVM learning algorithms, which was then applied to fast HEVC coding and fast H.264/AVC to

Manuscript received on Sep. 2016, revised Aug, 2017, accepted. Aug. 2017

This work was supported in part by the National Natural Science Foundation of China under Grant 61471348 and Grant 61501246, in part by Guangdong Natural Science Foundation for Distinguished Young Scholar under Grant 2016A030306022, in part by the National High-tech R&D Program of China under Grant 2015AA015901, in part by Shenzhen Science and Technology Development Project under Grant JSGG20160229202345378 and Shenzhen International Collaborative Research Project under Grant GJHZ20170314155404913, in part by the PhD Start-up Fund of Guangdong Natural Science Foundation under Grant 2015A030310262, and in part by Natural Science Foundation of Jiangsu Province under Grant BK20150930.

Y. Zhang and N. Li are with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China, (Email: {yun.zhang, na.li1}@siat.ac.cn)

Z. Pan is with School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China (Email: zqpan3-c@my.cityu.edu.hk);

X. Wang is with College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, 518060, China (Email: wangxu@szu.edu.cn)

G. Jiang is with Faculty of Information and Engineering, Ningbo University, Ningbo, 315211, China (Email: jianggangyi@nbu.edu.cn)

S. Kwong is with Department of Computer Science, City University of Hong Kong, Hong Kong, China (Email: cssamk@cityu.edu.hk)

Copyright © 2017 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology

HEVC transcoding, respectively. In [9], binary and multi-class SVMs were applied to INTER CU and PU size decisions in HEVC. Meanwhile, the multiple reviewers system was adopted to enhance the sample selection in the classifier learning. In addition, other learning algorithms including Linear Discriminant Functions (LDFs) [10], Markov Random Field (MRF) [11], and Decision Tree (DT) [12], *etc.*, were also investigated for the CU size decision. They were proposed for INTER frame coding where the temporal and motion information were adopted as key features. As compared with INTER coding, INTRA coding is more critical due to less available information and more complex CU partitions. Meanwhile, the overhead complexity of the learning algorithm is no more negligible, which makes the fast INTRA coding even more challenging.

INTRA coding in HEVC is very complex due to the adoption of quad-tree portioning structure and up to 35 prediction modes in each PU [13]. There are high demands of real-time and fast INTRA algorithms for HEVC, especially for the applications with limited computational resources such as ultra HD video on mobile platform. To reduce the complexity of CU size decision in INTRA coding, Shen *et al.* [14] proposed an early CU size decision algorithm based on texture homogeneity of the current CU and size of neighboring coded CU. Min *et al.* [15] proposed a fast INTRA CU size decision algorithm based on texture information, in which sub-CUs and CU's edge complexities in the horizontal, vertical, 45° and 135° diagonal directions were used as key features. Then, CU size was determined with hard thresholds for these edge complexities. Zhang *et al.* [16] proposed a fast INTRA CU size decision algorithm based on statistical modeling and correlation analyses, in which the most probable depth range was determined based on the spatial correlation among neighboring CUs, and a statistical model of RD cost was proposed to adaptively determine Early Termination (ET) thresholds varying with video contents and Quantization Parameters (QPs). In [17], CU depth and mode correlations between Enhancement Layer (EL) and Base Layer (BL) were exploited for fast INTRA coding in Scalable HEVC (SHVC). Moreover, CU modes of spatial neighboring CUs and co-located texture CU were used for the CU size decision in 3D-HEVC [18]. In addition to CU size decision, optimizations for INTRA angular prediction were also investigated in [19]-[21]. The texture edge detected by Sobel operator was utilized to predict the angular modes in each PU [19]. Based on the statistical properties of the neighboring reference samples, Wang *et al.* [20] checked modes from one of the three candidates (1, 19, and 35 INTRA modes) for each PU, so as to reduce the INTRA prediction complexity. In [21], two levels (CU and PU levels) of fast INTRA coding were proposed for HEVC. In CU size decision, CU depth can be early terminated when the aggregated RD cost of partial sub-CUs is larger than the current CUs. Meanwhile, Hadamard cost is used to reduce the INTRA angular mode candidates in Rough Mode Decision (RMD) at each PU level. Basically, they are statistical approaches in dealing with the prediction. However, due to the diversity of video content and increasing number of modes, it is difficult to discriminate the optimal mode from dozens of candidate modes with one or two features and thresholds. Meanwhile, it is more complicated if more features are included in traditional statistical approaches.

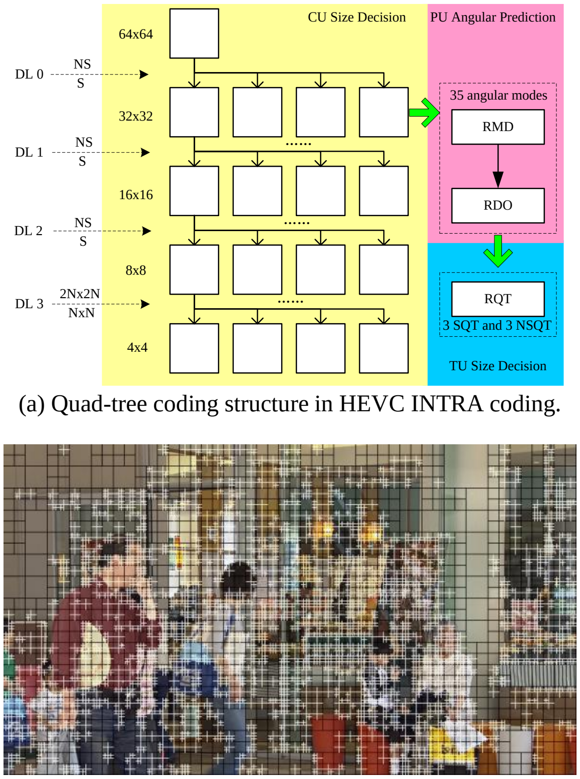
Machine learning is a hotspot in pattern recognition which learns from data and gives the optimal discriminative solution. With this advantage, a number of works [22]-[29] have been proposed for learning based INTRA mode decision in HEVC. In [22], fast PU skip and split termination algorithms were proposed for HEVC INTRA prediction. Firstly, the neighboring PU sizes were used to filter out PU sizes with low probabilities owing to the spatial correlation in a frame. In addition, Bayesian decision was used to determine whether to early terminate the PU coding process or not, in which Hadmard cost of the current CU divided by its neighboring Hadmard cost was used as a key feature. Similarly, Hu *et al.* [23] modeled the INTRA mode decision as Bayesian decision problems, in which Discrete Cosine Transform (DCT) coefficients and outlier information were utilized. In [24], progressive Bayesian classification composed of cascade online classifiers was proposed for INTRA CU size decision, in which Bayesian risk was used to balance the complexity and efficiency. However, online data collection and training will bring additional overhead. Random forests via offline training [25] and DT [26] were adopted to predict CU size of 64×64 or 32×32. In [27], DT classifiers were designed and adopted with chosen features that can distinguish types of blocks (natural image block or screen content block) for screen content coding. In [28], linear SVM classifiers with two features, *i.e.*, depth difference and Hadmard/RD costs among spatial neighboring CUs, were used to determine whether early terminate the CU decision or not. Meanwhile, gradients of the horizontal and vertical directions were used for the INTRA angular prediction in each CU. Additionally, a Convolution Neural Network (CNN) based fast algorithm was proposed to reduce the INTRA CU/PU modes for hardwired encoder [29]. However, CNN requires additional complexity and accelerators. Although many previous works have applied machine learning techniques for solving HEVC INTRA CU decision problems, the coding performance could be still further improved due to the following three reasons. Firstly, representative features need to be explored from the video content in order to improve the discriminability and reduce complexity overhead. Secondly, the optimal learning parameters for complexity allocation among different CU layers could be explored. Thirdly, ET and Early Skip (ES) strategies shall be jointly used for effective CU decisions.

TABLE I
RATIO OF CU DEPTH FOR DIFFERENT SEQUENCE AND QPS. [%].

QP	Sequence	D0	D1	D2	D3	D4
22	*BQSquare*	0.00	7.80	13.68	20.46	58.06
	BasketBallDrill	0.01	8.91	16.41	35.07	39.61
	FourPeople	1.62	23.15	36.45	23.36	15.42
	ParkScene	2.79	21.30	26.39	25.44	24.08
	Traffic	1.68	18.96	32.29	28.29	18.79
	Average	**1.22**	**16.03**	**25.04**	**26.52**	**31.19**
37	*BQSquare*	0.00	16.77	25.15	26.18	31.90
	BasketBallDrill	1.99	32.97	38.90	20.90	5.24
	FourPeople	11.00	33.08	34.61	17.53	3.78
	ParkScene	14.55	35.15	32.58	14.34	3.37
	Traffic	9.05	34.36	36.95	16.81	2.83
	Average	**7.32**	**30.47**	**33.64**	**19.15**	**9.42**

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology



(a) Quad-tree coding structure in HEVC INTRA coding.

(b) Example of block partition in *BQMall* (*QP* is 28).

Fig.1. CU partition in HEVC INTRA Coding.

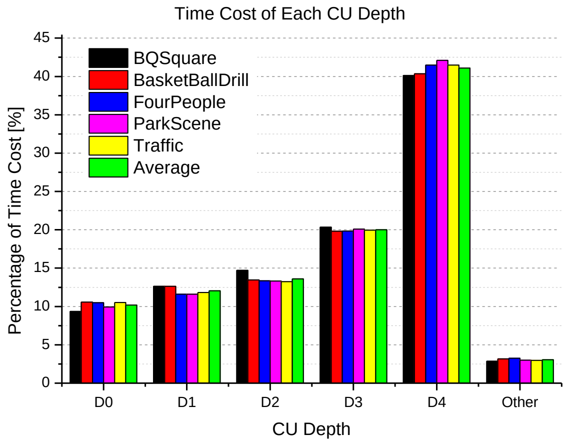


Fig. 2. Time cost of each CU depth.

TABLE II
COMPLEXITY REDUNDANCIES IN HEVC INTRA CODING. [%]

QP	22	27	32	37	**Average**
BQSquare	66.68	69.25	71.37	73.19	**70.12**
BasketBallDrill	70.57	75.21	78.85	80.97	**76.40**
FourPeople	78.00	79.35	80.69	82.09	**80.03**
ParkScene	75.51	77.92	80.50	82.86	**79.20**
Traffic	76.90	79.16	80.93	82.60	**79.90**
Average	**73.36**	**76.04**	**78.35**	**80.25**	**77.00**

In this paper, we propose effective data driven CU size decision approaches for HEVC INTRA coding. Two stages of classifications are proposed for early INTRA CU decision. Meanwhile, representative features of INTRA CU decision are explored and the optimal parameters derived from training data are achieved for complexity allocation at each CU layer. This paper is organized as follows, the statistical analyses of INTRA CU size and motivation are presented in Section II. The proposed CU size decision and optimization methodologies are described in Section III. Extensive experiments and analyses are performed in Section IV. Section V draws conclusions.

II. STATISTICAL ANALYSIS AND MOTIVATION

In HEVC, quad-tree coding structure of CU partition is adopted and there are four levels of CU depth, $i\in\{0,1,2,3\}$, where the CU size corresponds to 64×64, 32×32, 16×16 and 8×8, respectively, as shown in Fig.1 (a). The minimum CU size, *i.e.*, 8×8, can be further split into 4×4 block, which is conceptually the PU mode with SIZE_N×N. Fig.1(b) shows an example of CU partition for *BQMall* sequence coded with INTRA at Quantization Parameter (QP) equal to 28, where the b/w rectangles indicate CU/PU size partitions. For each CU/PU, the best mode will be selected from the 35 INTRA modes [13], which include 33 angular modes, 1 DC mode and 1 PLANAR mode. In the HEVC model, RMD is used to select a subset (3 to 10 modes) from 35 modes, which reduces the number of candidates for complex full RDO. Finally, in the Residual Quad-tree Transform (RQT), the optimal TU size shall be determined from up to three TU size candidates, *i.e.*, 32×32, 16×16 and 8×8. In HEVC INTRA coding, only symmetrical quad-tree transform (SQT) is activated. To sum up, there are three major loops of mode decisions, including CU size decision, PU INTRA prediction mode decision and TU size decision. They are basically determined by trying all loops and candidates, then selects the optimal one with the minimum RD cost, which is extremely complex. In this paper, we focus on the CU size decision, which is outer loop and possess major complexity part of the HEVC INTRA encoder.

We statistically analyze the CU depth distribution and time cost of each depth for different test sequences with various resolutions and contents, including *BQSquare* (416×240), *BasketballDrill* (832×480), *FourPeople* (1280×720), *ParkScene* (1920×1280) and *Traffic* (2560×1920). 100 frames of each sequence were encoded by HM 16.7 [31] with all INTRA main profile according to the Common Test Conditions (CTCs) [32]. *QP*s are 22, 27, 32, and 37. Table I shows CU depth distribution for different sequences and *QP* equals to 22 or 37, where D0 to D4 represent Depth 0 to Depth 4. We have the following three observations:

1) The distribution varies from sequences and *QPs*. Some sequences, *e.g.*, *BQSquare* and *BasketballDrill* are small size CU dominated, where the 8×8 and 4×4 CUs are higher than 75%. Whereas, some are larger size CU, *e.g.*, 32×32 and 16×16, dominated since they are with larger resolution and meanwhile more smooth regions are contained.
2) The average CU distribution of the five sequences from D0 to D4 are 1.22%, 16.03%, 25.04%, 26.52 %, and 31.19%, respectively, when *QP* is 22. The large size CUs, *e.g.*, 64×64, possess very small ratios. For small resolution sequences, such as *BQSquare* and *BasketballDrill*, there is no 64×64 CUs. Meanwhile, small size CUs (8×8 and 4×4) are dominated in the INTRA coding, which is different from the INTER coding [6] that large CU size dominated.

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology

3) The distributions are 7.32%, 30.47%, 33.64%, 19.15% and 9.42%, respectively when *QP* is 37. As the *QP* increases, the number of larger CUs increases and the number of smaller CUs decreases. The relative larger CUs (*e.g.* 32×32 and 16×16) are dominated. It is because while using larger *QP* to achieve higher compression ratio, the encoder intends to use larger size CUs leading to less bits but more distortions.

The Depth 0 and Depth 1 are with relative small ratio in INTRA coding, which is significantly different from that of INTER coding. Using ET scheme is more effective for the case that Depth 0 and Depth 1 are dominated or the probabilities of Depth 0 to Depth 4 decrease monotonically, but not effective in the case that small size CUs are dominated. Therefore, ES and ET shall be jointly utilized in fast HEVC INTRA coding.

In addition, we also analyze the coding time of each CU depth in HEVC INTRA coding. It is found from the Fig.2 that the time cost of each CU depth are almost consistent for different sequences and *QP*s. The average time cost of INTRA coding while using CU Depth 0 to 4 are 10.17%, 12.06%, 13.61%, 20.00 %, and 41.11%, respectively. Other coding overheads possess 3.05%. The complexity increases as the CU depth increases. Also, the time cost of CU Depth 3 and Depth 4 is much larger than those of Depth 0 to Depth 2, since there are more angular mode candidates in FULL RDO process in 8×8 and 4×4. Meanwhile, Discrete Sine Transform (DST) is additionally adopted in 4×4 TU decision. Table II shows the potential coding complexity redundancies in HEVC INTRA CU decision. These redundancies are calculated from CU distribution and CU time cost based on the assumption that the INTRA CU size can be 100% precisely predicted and unnecessary CU checking are all skipped. The complexity redundancies vary with sequences and *QP*s from 66.68% to 82.86% due to different CU distributions. Theoretically, there are 77.00% complexity on average can be reduced, which is the maximum complexity reduction of the proposed fast INTRA encoding algorithm.

III. The Proposed Data Driven CU Size Decision Approaches in HEVC INTRA Coding

A. Framework of the Proposed CU Size Decision

Compared with traditional statistical approaches, there are several unique advantages of the learning based scheme. Firstly, rather than traditional one or two hand-crafted key features, multiple features can be jointly used. Moreover, they can be effectively mapped to a non-linear high dimensional space, which improves the discriminative ability. Secondly, it learns a more accurate hyper-plane from sample data when compared with traditional statistical approaches. Thirdly, sophisticated training and learning algorithms can be adopted to achieve higher prediction accuracy. Since the CU size decision can be modeled as hierarchical binary classifications [6], we develop a data driven CU size decision for HEVC INTRA coding, which consists of two stages of classifications. Figure 3 shows the flowchart of the proposed CU size decision for depth *i*, which is a recursive process. For each input Coding Tree Unit (CTU), its features are input to a three-output classifier, which predicts the probable CU depth. Let NS, S and RDO in Fig.3 represent non-split, split and uncertain prediction that needs FULL RDO, respectively. If the classifier at the Decision Level (DL) *i* is the NS prediction, it only checks the current CU depth *i*. If it is the split prediction, it goes to the CU depth *i*+1 and checks the four sub-CUs at depth *i*+1. Since the classifier is with limited prediction accuracy and there are some feature spaces with high risk of false prediction, it goes to the FULL RDO prediction, in which the CU at the current depth *i* and four sub-CUs at the depth *i*+1 are both checked. Then, the optimal CU depth is determined based on the RD cost. Note that the proposed CU size decision for depth *i* is a recursive process and at the final CU depth, *i.e.*, when CU/PU size is 4×4, the proposed algorithm checks the SIZE_N×N mode. Therefore, the INTRA CU/PU size decision is designed from 64×64 to 4×4, and $i\in\{0,1,2,3,4\}$.

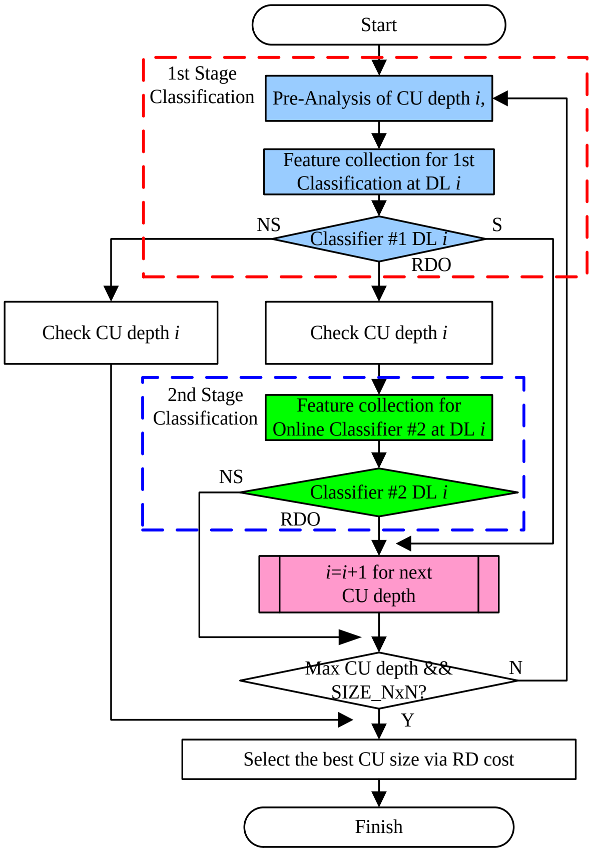


Fig. 3. Flowchart of the proposed CU size decision at CU depth *i*.

After the first stage classification/prediction, there are a number of samples predicted as uncertain. Therefore, for these samples, a second stage classification (see the blue dash rectangle in Fig. 3) is used to learn further from available information after checking CU depth *i*, such as bit and RD cost of the current CU. Based on these collected new feature values and online learning, the second stage classification is used to determine the uncertain samples and decide whether the current CU is non-spilt and early terminate the CU decision process or not. These two stage classifications are effective ways of exploiting the available information and meanwhile maintain low complexity.

In addition, at the first stage CU prediction, offline training is adopted since it has the following advantages: larger number and more diverse training samples could be included since the

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology

complexity overhead of training and training sample collection will not be included in the encoder; meanwhile, different *QP*s and video contents could be used in the model training. In the second stage classification, online model training is adopted since it can further exploit the correlation between the training and predicting sets. Meanwhile, it is complementary to the offline training models and is useful to further discriminate the uncertain samples from the first stage decision. The integration of online and offline training can more effectively exploit the advantages of the two different learning ways.

B. Features Extraction and Selection for the CU Size Decision

Based on the features analyses in fast INTRA CU size decision schemes [14]-[28], four categories of factors are highly correlated to CU size and shall be adopted as the key features in INTRA coding. The first category is texture information of the video contents since smooth region usually leads to large CU size. The second category is pre-analysis information of the current CU before coding, which represents video content information. The third category is the context information of the current CU since there are spatial correlation among the neighboring CUs. The last category is the coding information of the current CU, which gives more details and accurate coding information. Details of the representative features are presented as follows:

1) **Texture information:** It is of higher efficiency by using smaller CU size to encode the texture regions, whereas using larger CU size to encode the smooth regions. Based on this property, texture information (denoted as $x_T(i)$) will be one kind of useful information for CU size decision. Here, we calculate the texture of each CU/CTU based on the Mean Absolute Difference (MAD) of luminance component, which is

$$T(\mathbf{B})=\frac{1}{N_B}\sum_{I(i,j)\in\mathbf{B}}\left|I(i,j)-\frac{1}{N_B}\sum_{I(i,j)\in\mathbf{B}}I(i,j)\right|, \qquad (1)$$

where **B** is the block of a CU or CTU, N_B is the number of pixels in block **B**. $I(i,j)$ is the luminance value of pixel located at (i,j) in block **B**. In addition, the texture difference between the current CU (**B**) and its corresponding four sub-CUs ($\mathbf{b}_i$) are also adopted as a feature (denoted as $x_{TD}(i)$), which is calculated as

$$V(\mathbf{B})=T(\mathbf{B})-\sum_{i=0}^{3}T(\mathbf{b}_i), \qquad (2)$$

For smooth regions that have high probabilities of using large CU size, the texture difference value is small. On contrary, this difference value becomes larger for texture regions.

2) **Pre-analysis of the current CU:** To exploit the information of the current CU more precisely, we pre-analyze the current CU by encoding it with PLANAR mode, and calculates its RD cost and distortion. The first feature is the RD cost with the PLANAR mode and normalized by *Qstep* (*i.e.*, divided by *Qstep*). It is denoted as $x_{RD/Q}(i)$. The second feature is the RD cost divided by distortion, denoted by as $x_{RD/D}(i)$. Here, we calculated the RD cost only using PLANAR mode has two main reasons. Firstly, the complexity of the pre-analysis process must be low, since it is an overhead complexity of the proposed algorithm. Using PLANAR mode only can avoid the complex angular prediction mode decision and has much lower complexity. Secondly, the PLANAR mode is of the highest probability to be selected as the best mode in the angular mode decision. Thus, its RD cost is of the most probably being close to the real RD cost via full angular mode decision under the circumstance that we only check one mode candidate.

3) **Context information:** Since there is a high spatial correlation inside of an image, the context information from neighboring CU or CTUs shall be used for CU size decision of the current CU. For example, if the neighboring CUs are both encoded with small CU size, *i.e.*, 8×8 or 4×4, the optimal CU size of the current CU is probably small CU size, *i.e.*, 8×8 or 4×4. In this paper, three types of context information are used, including the neighboring RD cost, CU depth and PU mode of neighboring CTUs or CUs. The first is the average RD cost of the left and above neighboring CTUs, denoted as $x_{NB\_CTU\_RD}$.

The second is the sum of CU depth of the neighboring CTUs, denoted as $x_{NB\_CTU\_Depth.}$ In HEVC, the CU depth of each CTU are labeled with 4×4 unit, *i.e.*, a 64×64 CTU consists of 256 4×4 units. Therefore, each CTU has 256 labels of CU depth, each label value is from 0 to 3. It is calculated as

$$x_{NB\_CTU\_Depth}=\sum_{k=0}^{255}\left[D_{L,4\times4}(k)+D_{\mathrm{A},4\times4}(k)\right], \qquad (3)$$

where $D_{L,4\times4}(k)$ and $D_{A,4\times4}(k)$ represent the CU depth in 4×4 unit at position k in left and above CTU, respectively. For example, for a CTU partitioned with four 32×32, the 256 labels are all 1 (CU depth is 1) and the $x_{NB\_CTU\_Depth}$ is 256. If the CTU is partitioned with 8×8, the 256 labels are all 3 (CU depth is 3) and the $x_{NB\_CTU\_Depth}$ is 256×3. This value indicates the CU partitions of the neighboring CTUs. If this value is large, there is a higher probability using smaller size CU for the current CU/CTU due to spatial texture correlation of the video content.

The third is CU/PU depth of the left and above neighboring CUs, denoted as $x_{NB\_CU+PU\_Depth}(i)$. In HEVC standard, the CU Depth 0 to 3 are corresponding to CU size 64×64 to 8×8. Then, PU modes (SIZE_N×N, SIZE_2N×2N) are used to indicate whether an 8×8 CU is split into four 4×4 blocks or not. Thus, the sum of neighbor CU and PU depth value is calculated as

$$x_{NB\_CU+PU\_Depth}(i)=\left[D_L(i)+P_L(i)+D_A(i)+P_A(i)\right]/2, \qquad (4)$$

where $D_L(i)$ and $D_A(i)$ are CU depth of the left and above neighboring CUs if available. $P_L(i)$ and $P_A(i)$ are the PU mode of left and above CUs, which is 0 for SIZE_2N×2N and 1 for SIZE_N×N. The features $x_{NB\_CTU\_Depth}$ and $x_{NB\_CTU\_RD}$ are for a CTU and usually have less correlation with the small size CUs. In other words, neighboring CTU context information are more effective in determining larger CUs rather than the smaller CUs. Therefore, $x_{NB\_CTU\_Depth}$ and x_{CTU_RD} are adopted in DL 0 and 1, while $x_{NB\_CU+PU\_Depth}$ is adopted in DL 2 and 3.

4) **Coding information of the current CU.** The coding information of CU depth *i* includes the RD cost and coding bits, denoted as $x_{RD}(i)$ and $x_{Bit}(i)$, respectively. These two information are more accurate than the RD cost from the pre-analysis in representing the current CU. Basically, as the RD cost and coding bits are small, there has a higher probability of selecting non-split mode as the best mode. Since it is only available after checking the current CU depth, these information are used in the second stage classification and the online trained classifier.

Features selection in this paper considers the importance of features, number of features, prediction accuracy and computational complexity. Representative features are

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology

TABLE III
SELECTED FEATURES FOR THE TWO STAGES OF CLASSIFIERS.

No.	Category	Features	1st Stage Classification				2nd Stage Classification			
			DL0	DL1	DL2	DL3	DL0	DL1	DL2	DL3
1	Texture Information	$x_T(i)$	√	√	√	√	√	√	√	√
		$x_{TD}(i)$	√	√						
2	Pre-analysis of Current CU	$x_{RD/Q}(i)$	√	√	√	√				
		$x_{RD/D}(i)$	√	√						
3	Context Information	$x_{NB\_CTU\_RD}$	√	√						
		$x_{NB\_CTU\_Depth}$	√	√			√	√		
		$x_{NB\_CU+PU\_Depth}(i)$			√	√			√	√
4	Coding Information of Current CU	$x_{RD}(i)$					√	√	√	√
		$x_{Bit}(i)$					√	√	√	√

extracted as the above mentioned. More features included in the classifications will bring higher prediction accuracy. However, they also lead to more extraction time and increase the dimension of feature vector, which consequently increase complexity overhead. In addition, the coding information of the current CU is only available after checking the current CU and therefore can only be used in the second stage classification. Table III shows the selected features for the two stages classifications, where symbol "√" indicates the selected feature at each DL. Since the complexity is critical in HEVC INTRA coding, the number of features shall be minimized subject to high prediction accuracy. Therefore, in the DL0 and DL1 of the first stage classification, six features including texture ($x_T(i)$, $x_{TD}(i)$), pre-analysis ($x_{RD/Q}(i)$, $x_{RD/D}(i)$) and context ($x_{NB\_CTU\_RD}$, $x_{NB\_CTU\_Depth}$) are adopted. Since the classifiers in DL2 and DL3 are used more frequently, the complexity overhead is more critical, we only adopt one from each category, which are $x_T(i)$, $x_{RD/Q}(i)$ and $x_{NB\_CU+PU\_Depth}(i)$.

In addition, four features are adopted for the second stage classification which is implemented after checking the current CU depth. At this stage, coding information of the current CU is now available. Therefore, $x_T(i)$, $x_{RD}(i)$ and $x_{Bit}(i)$ are adopted. Basically, when $x_{RD}(i)$ and $x_{Bit}(i)$ are small, the current CU depth is of high probability of being selected as the best CU and the rest CU depth is not necessary to be checked. As for the feature of context information, $x_{NB\_CTU\_Depth}$ is used for DL0 and DL1 while $x_{NB\_CU+PU\_Depth}(i)$ is used for DL2 and DL3. This is because $x_{NB\_CTU\_Depth}$ is from neighboring CTUs and has higher correlation to the large CUs. For the small CUs, $x_{NB\_CTU\_Depth}$ is distant and the neighbor CU information $x_{NB\_CU+PU\_Depth}$ has closer relation with the current CU size.

C. Sample Selection for CU Size Decision

In addition to feature extraction and selection, sample selection plays an important role in classification model training. The first principle of the sample selection in this work is the trade-off between prediction accuracy of the trained model and the complexity overhead of the SVM prediction. In the first stage offline training process, we select five sequences with different image contents and resolutions, which are *BQSquare* (416×240), *BasketballDrill* (832×480), *FourPeople* (1280× 720), *ParkScene* (1920×1080) and *Traffic* (2560×1600). These sequences were encoded with *QP*s 22, 27, 32 and 37, respectively. Then, the features and the split/non-split labels were collected for the model training. We did not use many frames for each sequence, because the computational complexity of the SVM prediction is direct proportional to the number of supporting vectors, which increases dramatically along with the number of training samples. Therefore, to achieve a good trade-off between the sample variety (different video contents, resolutions and *QP*s) and the complexity of the SVM prediction, we limit the number of training samples and only one frame of each sequence (encoded with four *QP*s) are collected. In the second stage classification, N_{train} frames are encoded while the first stage classification is activated, then feature values and coding modes are collected as ground truth for the online training. The next N_{pred} frames are encoded with the second stage classification, which predicts whether early terminate the sub-CU checking or not. N_{train}+N_{pred} is a cycle of learning and predicting of the second stage classification. Large N_{train} leads to more samples for robust training, however, it increases the complexity overhead since it increases the time cost of data collection, training and prediction as well.

The second important principle of the sample selection is to keep a balance between the number of positive and negative samples. In this work, we randomly dropped the positive samples if they are much more than the number of negative samples, and vice versa. Then, the ratio of positive to negative samples is close to 1:1 for the first stage of classification. For the second stage classification, since it is an online learning and difficult to predict the ratio of positive to negative samples while encoding various video contents, we did not dropped the positive or negative samples in the online training.

Thirdly, the distributions of the CU depths is uneven and varies with video contents and *QP* values, as shown in Table I. Usually, there are large number of small size CUs and small number of large size CUs (*e.g.*, 64×64) in INTRA coding. Since the number of training frames is set, we randomly dropped small size CUs (samples) so as to reduce the total number of training samples for classifiers at larger DLs.

D. SVM based Classifier Design for the CU Decision

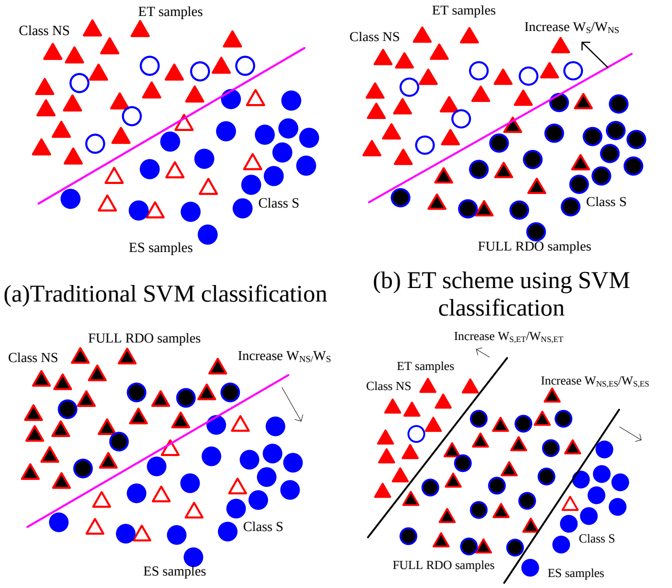
Since SVM has good performance in tackling the numerical problems, it is adopted as the key classifier. Let $\mathbf{x}_i$, i=1,2,3…., N be the feature vectors of the training set, $\mathbf{X}$. The i^{th} sample in the set can be represented as {$\mathbf{x}_i$, y_i}, where y_i is the class label, +1 for class S (split) and -1 for class NS (non-split). Samples are separable by a designed hyper-plane, which is

$$f(\mathbf{x}) = \boldsymbol{\omega}^T \phi(\mathbf{x}) + \omega_0 , \tag{5}$$

where $\phi(\cdot)$ is non-linear mapping function for the feature spaces. $\boldsymbol{\omega}$ is a trained model. To achieve an optimal hyper-plane to

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology



(a)Traditional SVM classification

(b) ET scheme using SVM classification

(c) ES scheme using SVM classification

(d) Joint ET and ES scheme using three-output classification

Fig.4. Examples SVM classification based ET and ES scheme.

separate the samples, cost function is determined as

$$J\left(\boldsymbol{\omega},\omega_0,\boldsymbol{\xi}\right)=\frac{1}{2}\left\|\boldsymbol{\omega}\right\|^2+CW_{NS}\sum_{i=1}^{N_{NS}}I\left(\xi_i\right)+CW_S\sum_{i=1}^{N_S}I\left(\xi_i\right), \qquad (6)$$

$$s.t.\ y_i\left[\boldsymbol{\omega}^T\mathbf{x}_i+\omega_0\right]\ge 1-\xi_i,\xi_i\ge 0,i=1,2,\ldots N$$

where C is a positive constant controlling the relative influence of the two competing terms. $\boldsymbol{\xi}$ is a vector of ξ_i. W_{NS} and W_S are weighted factors for class NS and S that can adjust the hyper-plane and the prediction accuracy. N_S and N_{NS} are numbers of S and NS class samples, $N= N_S+N_{NS}$. In this paper, Radial Basis Function (RBF) is adopted as the kernel function for the SVM, since it handles the non-linear case well and has fewer numerical difficulties [30] when the number of features is small. In addition, the constant C is to have a trade-off between training error and the flatness of the solution. The larger value of C is the less the final training error will be. Accordingly, it is set as 100 [6] in this paper, which gives relative larger penalty to misclassification as compared with the default value.

Fig. 4 (a) is the hyper-plane determined by using the same weighted factors, *i.e.*, W_{NS} and W_S of the classifier are both set as 1. Let N_{FT} and N_{UT} be the number of filled and unfilled triangles, *i.e.*, ▲ and △, and let N_{FC} and N_{UC} be the number of filled and unfilled circuits, *i.e.*, ● and ○. Let N_{ALL} be the total number of samples and $N_{ALL} = N_{FT}+ N_{UT}+N_{FC}+N_{UC}$. In this case, samples (▲ and ○) are classified as NS mode and early terminate the rest CU coding process. Thus, the ET rate is $(N_{FT}+N_{UC})/N_{ALL}$. Samples (●+△) are classified as S and ES scheme is activated for the current CU depth. Thus, the ES rate is $(N_{FC}+N_{UT})/N_{ALL}$. The misclassified samples (○+△) will lead to the RD degradation. Due to the variety of video contents and high similarities between CU modes in upper and lower layers, it is difficult to find a powerful learning algorithm and representative features that can always maintain sufficient prediction accuracy, *e.g.*, $(N_{FT}+N_{FC})/N_{ALL}$ is higher than 95% or $(N_{UC}+N_{UT})/N_{ALL}$ is smaller than 5%. Some positive and negative samples may stick together and difficult to be distinguished, *i.e.*, non-separable samples. Fig. 4 (a) shows a classification illustration of an optimal model learned from given features. However, there are also many misclassified samples, shown as the unfilled circles and triangles.

Fig.4 (b) and (c) shows the testing examples of ET and ES schemes using the SVM classification, respectively, where the black filled samples (▲+●) are checked with FULL RDO, (▲+○) are ET samples, and (●+△) are ES samples. For the split (S) prediction, the current CU depth can be skipped and goes to next CU depth directly, which is ES process. For the NS prediction, the current CU depth is checked and the rest CU depths are skipped, which is called ET. For the ET scheme in Fig. 4 (b), samples (▲+○) classified as NS will be early terminated after checking the current CU depth. Meanwhile, black filled samples (▲+●) are FULL RDO samples that encoded by the original FULL RDO process, in which neither ES nor ET is used. FULL RDO samples can 100% accurately obtain the best mode since the coding process is identical to the original HEVC encoder. As we use the C-SVM classifier, the optimal hyper-plane can be adjusted by giving different penalties to the misclassification of class NS or S. For example, increasing W_S/W_{NS} gives higher penalty to misclassification of Class S. The numbers of ET samples and the misclassified samples (○) will be reduced at the same time. Meanwhile, FULL RDO samples are increased, which indicates less complexity reduction. Similarly, Fig. 4 (c) is an illustration of ES scheme. Samples (●+△) are classified as S and checking the current CU depth will be skipped. As W_{NS}/W_S increases, the hyper-plane goes down and ES rate reduces. Meanwhile, the number of misclassified samples (△) will be reduced and the number of FULL RDO samples will be increased.

To reduce the number of misclassified samples (both NS and S), we can adjust the weighted factors and combine Fig. 4 (b) and (c) to be Fig. 4 (d), which is adopted from [6]. Two C-SVM classifiers with different parameters (W_S and W_{NS}) are adopted. The weights (W_S and W_{NS}) for the ET classifier are denoted as ($W_{S,ET}$, $W_{NS,ET}$). Meanwhile, the weights (W_S and W_{NS}) for the ES classifier are denoted as ($W_{S,ES}$, $W_{NS,ES}$). As we increase the weight $W_{S,ET}/W_{NS,ET}$ for Classifier #1 and meanwhile decrease the weight $W_{S,ES}/W_{NS,ES}$ for the other classifier, Classifier #2, the number of samples classified as FULL RDO are increased in both up and down directions, as shown in Fig. 4 (d). The output of the new designed three-output classifier (O_{C1}) is defined as

$$O_{C1}=\begin{cases} NS & O_1=NS\ \&\ O_2=NS \\ RDO & O_1\ne O_2 \\ S & O_1=S\ \&\ O_2=S \end{cases}, \qquad (7)$$

where O_1 and O_2 are the outputs of classifier #1 and #2, which are S (+1) or NS (-1). RDO indicates FULL RDO prediction. The CU size of these FULL RDO samples can be 100% accurately determined, *i.e.*, no RD loss. However, no complexity reduction could be achieved since both split and non-split are checked, which is no gain either. Overall, the number of misclassified NS samples (red unfilled triangles △) and S samples (blue circles ○) are both reduced significantly.

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology

In the proposed algorithm (see Fig.3), the three-output classification (Fig. 4 (d)) is used for the first stage classification. Meanwhile, to further ET the FULL RDO samples, an ET scheme (Fig. 4 (b)) is used for the second stage classification. The output of the second stage classifier (O_{C2}) is presented as

$$O_{C2} = \begin{cases} NS & O_1 = NS \\ RDO & O_1 = S \end{cases}, \tag{8}$$

where O_1 is the output of a binary SVM classifier.

IV. Optimal Parameter Determination for Two Stages of Learning Classifiers

A. *The Optimal Complexity Allocation among DLs for the First Stage Classification*

False predictions at different DLs may lead to different RD cost increase; meanwhile, the RD cost increases caused by false positive and false negative prediction are also different. To obtain the optimal weighted factors for the classifiers #1 and #2 (corresponding to ET and ES schemes), we analyze the impacts of the miss-classification on the RD cost and complexity reduction in each CU DLs. There are two schemes in this analytical experiment. One is the ET scheme in which the CU is predicted as non-split (Class NS) and the rest CU level is not necessary to be checked, *i.e.*, early terminate the coding process. The other is ES scheme in which the CU is predicted as spilt (Class S) and the current CU coding process is skipped. The two schemes corresponds to Fig.4 (b) and (c).

To quantify the impacts of the weighted factors (W_S and W_{NS}) in the ET and ES schemes at each CU decision layer from DL0 to DL3, five test video sequences with different resolutions and contents were encoded, which are *BlowingBubbles* (416×240), *PartyScene* (832×480), *Vidyo3* (1280×720), *BasketballDrive* (1920×1080) and *Traffic* (2560×1600). *QP* is 22, 27, 32 and 37. While training models for the first stage classification, the ratio W_S:W_{NS} are configured from 5:1 to 1:5 with 0.5 step, *i.e.*, {5:1, 4.5:1, 4:1, …1:1, 1:1.5, 1:2,….1:4.5, 1:5}.

Fig.5 shows the Bjonteggard Delta Bit Rate (BDBR) and complexity reduction for ET scheme at DL *i*, which are with legend "DL*i*_ET", $i\in\{0,1,2,3\}$. The *y*-axis is BDBR of the ET or ES scheme when comparing to the original HM and the *x*-axis is the parameter W_S/W_{NS} in log scale, *i.e.*, $\text{Log}_2(W_S/W_{NS})$. We can observe that the BDBR increases from 0% to 15% for DL0 to DL3 as $\text{Log}_2(W_S/W_{NS})$ increases. The BDBR is less than 0.5% when $\text{Log}_2(W_S/W_{NS})$ is smaller than -2.0. As the ET schemes at DL0 to DL3 are simultaneously applied to the final optimized scheme, the BDBRs and complexity reductions are additive. Therefore, as we shall maintain the overall BDBR is smaller than a target bit rate (ΔR_T), small values of $\text{Log}_2(W_S/W_{NS})$ shall be selected so as to constrain the BDBR at each DL within a sufficient low value. Meanwhile, Fig.5 also shows the BDBR for ES scheme, which are with legend "DL*i*_ES", $i\in\{0,1,2,3\}$. The BDBR values decrease as $\text{Log}_2(W_S/W_{NS})$ increases. The BDBR value is very small for all $\text{Log}_2(W_S/W_{NS})$ at DL0 which indicates the learning algorithm is essentially accurate for the ES scheme at DL0. For DL1 to DL3, we shall use large $\text{Log}_2(W_S/W_{NS})$ in order to guarantee the overall BDBR from the four DLs is negligible. The BDBR increase (ΔR_m) of the ET and ES schemes at the four DLs are fitted with Logistic function, which can be presented as

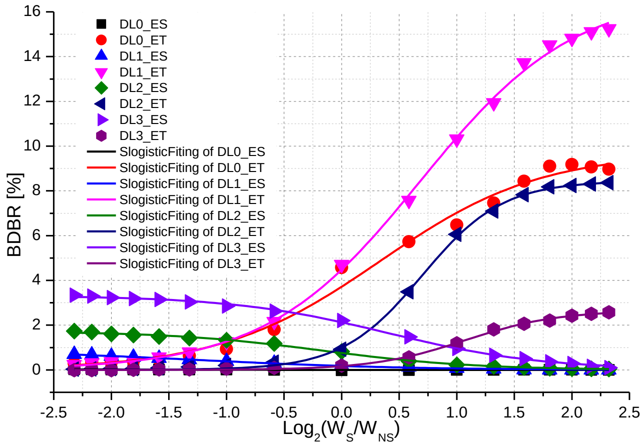


Fig.5. BDBR for the ES and ET schemes with different weighted factors.

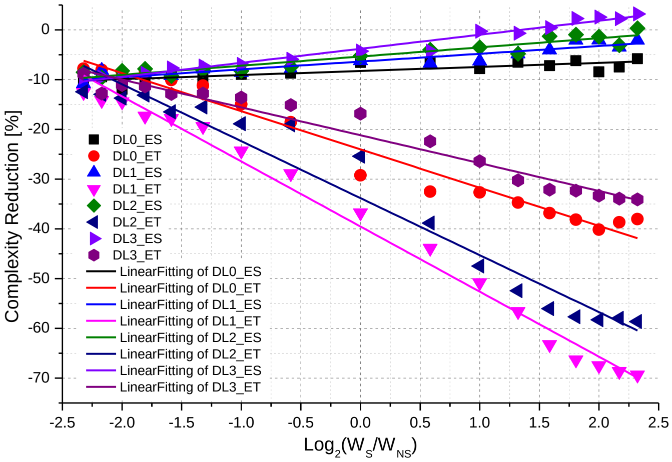


Fig.6. Complexity reduction for the ES and ET schemes with different weighted factors.

$$\Delta R_m = \frac{p_m}{1+e^{-k_m\left(x_m - x_{c,m}\right)}}, \tag{9}$$

where k_m, p_m and $x_{c,m}$ are fitting parameters, x_m is $\text{Log}_2(W_S/W_{NS})$ of the scheme *m*, *m* represents ET and ES schemes at the four different DLs, $m\in\{\text{DL}i\_\text{ES}\}\cup\{\text{DL}i\_\text{ET}\}$, $i\in\{0,1,2,3\}$

Fig.6 shows the complexity reduction by individual ES and ET scheme at each DL, where the dots are the real collected data under different W_S/W_{NS} configurations and the curves are the linear fitted results. The *y*-axis is the complexity reduction and the *x*-axis shows different parameters, *i.e.*, $\text{Log}_2(W_S/W_{NS})$. It is found that the complexity reduction increases as $\text{Log}_2(W_S/W_{NS})$ increases for the ET schemes, and meanwhile, it decreases as $\text{Log}_2(W_S/W_{NS})$ increases for the ES schemes. For the DL3_ES scheme, the complexity reduction is positive because the overhead caused by the feature extraction and prediction using the classifier is more than complexity reduction from ES. The complexity reduction of ES and ET scheme at each DL can be fitted with linear function as

$$\Delta T_m = a_m + b_m x_m, \tag{10}$$

where a_m and b_m are the fitting parameters for the scheme *m*. The detailed fitting parameters in Eqs. (9) and (10) and fitting accuracy are shown in Table IV.

According to the complexity reduction and BDBR increases, our optimization objective is finding optimal vector $\mathbf{X}=\{x_m\}$ to

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology

TABLE IV
FITTING PARAMETERS, ACCURACY AND THE OPTIMAL WEIGHTED FACTOR SETS.

		ES scheme				ET scheme			
m		DL0	DL1	DL2	DL3	DL0	DL1	DL2	DL3
Logistic Fitting	p_m	0.023	0.888	1.729	3.321	9.730	16.860	8.443	2.661
	k_m	-2.942	-1.113	-1.507	-1.480	1.434	1.490	2.870	2.482
	$x_{c,m}$	-1.285	-1.242	-0.165	0.393	0.336	0.683	0.700	1.076
	Accuracy (R^2)	0.993	0.995	0.997	0.998	0.992	0.999	0.999	0.999
Linear Fitting	a_m	-8.262	-6.368	-5.357	-3.796	-24.033	-39.530	-33.829	-21.185
	b_m	0.820	1.565	1.844	2.821	-7.688	-13.088	-11.457	-5.603
	Accuracy (R^2)	0.859	0.889	0.895	0.973	0.968	0.988	0.956	0.956
ΔR_T=1.0%	x_m=Log$_2$(W_S/W_{NS})	-1.287	1.112	2.158	2.906	-2.287	-1.881	-0.666	-0.255
ΔR_T=1.0%	(W_S,W_{NS})	(1:2.440)	(2.161:1)	(4.464:1)	(7.497:1)	(1:4.881)	(1:3.683)	(1:1.587)	(1:1.193)
ΔR_T=1.5%	x_m=Log$_2$(W_S/W_{NS})	-1.288	0.726	1.906	2.655	-2.030	-1.634	-0.539	-0.101
ΔR_T=1.5%	(W_S,W_{NS})	(1:2.442)	(1.654:1)	(3.748:1)	(6.298:1)	(1:4.084)	(1:3.104)	(1:1.453)	(1:1.073)

TABLE V.
RATIOS OF SPLIT AND NON-SPLIT PREDICTIONS AND THEIR PREDICTION ACCURACIES [%].

Sequences	DL i	$P_S(i)$	$P_{NS}(i)$	$P_{RDO}(i)$	$A_S(i)$	$A_{NS}(i)$
BlowingBubb.	0	100.00	0.00	0.00	100.00	INVALID
	1	76.72	1.93	21.35	100.00	86.40
	2	63.85	3.40	32.75	99.93	86.79
	3	63.47	7.10	29.43	97.48	100.00
PartyScene	0	100.00	0.00	0.00	100.00	INVALID
	1	88.26	0.81	10.93	100.00	100.00
	2	82.23	1.66	16.11	99.99	99.19
	3	78.04	3.30	18.66	97.46	100.00
BasketballDr.	0	64.90	26.74	8.36	100.00	100.00
	1	16.19	38.21	45.60	100.00	100.00
	2	20.49	19.82	59.70	99.30	99.49
	3	31.07	21.85	47.08	93.93	100.00
Vidyo3	0	80.73	13.31	5.96	100.00	100.00
	1	24.22	25.08	50.70	100.00	100.00
	2	21.54	31.59	46.87	98.17	98.69
	3	35.92	20.58	43.50	80.15	100.00
Traffic	0	90.40	5.57	4.03	100.00	100.00
	1	30.66	14.11	55.23	100.00	100.00
	2	30.40	17.03	52.58	99.51	99.64
	3	35.88	16.28	47.84	94.82	100.00
Average	**0**	**87.21**	**9.12**	**3.67**	**100.00**	**100.00**
	1	**47.21**	**16.03**	**36.76**	**100.00**	**97.28**
	2	**43.70**	**14.70**	**41.60**	**99.38**	**96.76**
	3	**48.88**	**13.82**	**37.30**	**92.77**	**100.00**

*Note that the INVALID indicates there is no NS prediction and the accuracy is invalid. They are not taken into account in calculating the average values

minimize the total computational complexity subject to the total BDBR increase is smaller than the target ΔR_T, which is presented as

$$\min_{\mathbf{x}}\left(1-\sum\Delta T_m\right), s.t.\sum\Delta R_m \le \Delta R_T . \qquad (11)$$

Minimize the computational complexity of the encoder is to maximize the complexity reduction of the proposed algorithm. Thus, Eq. (11) can be solved by transforming it to be a Largrange optimization function, which is presented as

$$\mathbf{X}=\arg\max_{\mathbf{x}} J,\ J=\sum\Delta T_m+\lambda\left(\sum\Delta R_m-\Delta R_T\right). \qquad (12)$$

It is easy to prove Eq. (12) is convex. Thus, we take partial derivation to each x_m and λ, and then set them to 0. Solve these nine equations, the optimal x_m can be obtained for each ET and ES schemes. Table IV shows the fitting parameters, accuracy and the optimal weighted factor sets for training the offline classifiers. The bottom four rows of Table IV show the optimal x_m and corresponding (W_S,W_{NS}) for the ET and ES schemes at four DLs given the target ΔR_T with 1% or 1.5%, respectively. Given the optimal x_i, its corresponding (W_S,W_{NS}) can be determined. For example, at DL1@ΔR_T=1.0%, x_m is 1.112 and -1.881 for the ES and ET schemes, respectively. Then, the corresponding (W_S,W_{NS}) to the x_m will be set as (2.161:1) and (1:3.683). The (W_S,W_{NS}) for training the classifiers at the rest DLs can be derived accordingly, as shown in Table IV. Similarly, x_i and (W_S,W_{NS}) can be calculated by using the parameters in Table IV if we set another target ΔR_T, *e.g.* 2.0%.

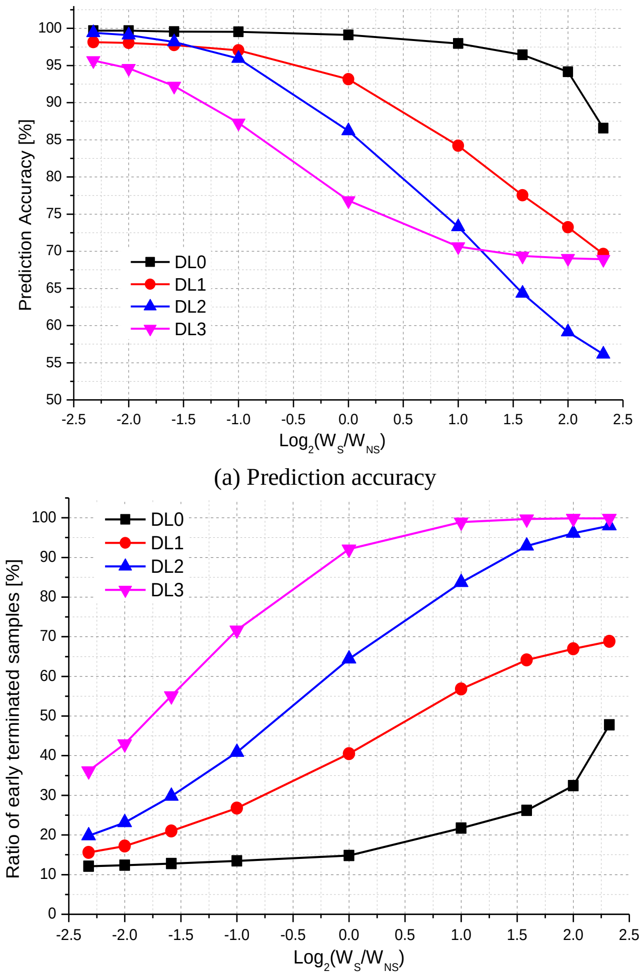
To further testify the performance of the first stage classification while using the obtained optimal weighted factors (W_S,W_{NS}), we count the ratios of split and non-split prediction as well as their prediction accuracies, as shown in Table V. $P_S(i)$, $P_{NS}(i)$ and $P_{RDO}(i)$ represent the ratios of split, non-split and FULL RDO prediction at the DL i while the first stage of classifications are applied in the HEVC CU decision. Last two columns, $A_S(i)$ and $A_{NS}(i)$ represent the accuracies of split and non-split prediction in the INTRA coding. We observe that the average $A_S(i)$ is 100%, 100%, 99.38% and 92.77%, while the average $A_{NS}(i)$ is 100%, 97.28%, 96.76% and 100%, respectively, from DL0 to DL3, which indicates the prediction is highly accurate. On the other hand, $P_S(i)$, $P_{NS}(i)$ and $P_{RDO}(i)$ varies with video contents, DLs and *QP*s according to the test results. $P_S(i)$ is 87.21%, 47.21%, 43.70% and 48.88%, respectively. $P_{NS}(i)$ is 9.12%, 16.03%, 14.70% and 13.82%, respectively. $P_S(i)$ is dominated as compared with $P_{NS}(i)$, which is reasonable since the small size CU is dominated according to the CU distribution statistics in Table I. Basically, it is found that the proposed classifiers are effective and have sufficient high prediction accuracy. Meanwhile, the classification can adapt to various video contents, DLs and *QP*s and work well.

B. The Optimal Parameter Determination for the Second Stage Classification.

In the first stage, the feature information input to the learning algorithm includes pre-analysis information, texture information and previously coded context information. However, these information are obtained before encoding the current CU depth. Due to limited feature information at the first stage, *i.e.*, without the encoding information of the current CU, the prediction accuracy of the learning algorithm is limited. We thus may need to increase the number of FULL RDO samples to improve the prediction accuracy of the first stage classification. As tabulated in Table V, it is observed that the

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology



(a) Prediction accuracy

(b) Ratio of early terminated samples

Fig.7. Prediction accuracy and ratio of early terminated samples for the second stage classification.

ratio of FULL RDO samples ($P_{RDO}(i)$) are relative small for *PartyScene*, which are 0%, 10.93%, 16.11% and 18.66% respectively from DL0 to DL3. However, they are large for *BasketballDrill*, *Vidyo3* and *Traffic* sequences, the $P_{\mathrm{RDO}}(i)$ value of which is up to 59.70%. On the other hand, the $P_{\mathrm{RDO}}(i)$ is from 0% to 8.36% for DL0 and increases for DL1 to DL3. Overall, the average ratio of FULL RDO samples are 3.67%, 36.76%, 41.60%, and 37.30%, respectively, from DL0 to DL3. Since the FULL RDO sample is relative large and shall be further early terminated for more complexity reduction, we thus propose the second stage classification and the corresponding ET algorithm. In this stage, classification is adopted to determine whether skip checking the rest CU depth or not. Therefore, only one binary SVM classifier is required, in which the NS is skip the sub-CU and RDO checks the sub-CU, shown as Fig. 4(b) and Eq.(8).

Since the current CU depth has been encoded and checked in the first stage, the by-product information of encoding the current CU depth is available for the second stage in determining whether to early terminate the rest CU depth. Here, the number of features is strictly controlled since the prediction complexity of the SVM is in direct proportional to the number of features as well as supporting vectors. In this stage, four dimensional feature vectors are adopted, including the RD cost of the current CU depth (x_{RD}), the number of bit of the current CU depth (x_{bit}), and texture information (x_T). The fourth feature is $x_{NB\_CTU\_Depth}$ when the current CU depth is small than 2. Otherwise, the fourth feature is $x_{NB\_CU+PU\_Depth}$. Details of the adopted features for online classifiers are shown in Table III.

Models of the first stage classification are from an offline training scheme. The advantages of the offline training include 1) large training set with various content and properties. Also, sophisticated training process could be included and it is not required in the encoding process. 2) the training samples and ground truth do not need to be collected in the coding. However, the disadvantage is that the training samples may significantly different from the samples and parameters (*e.g.*, *QP*s) in the video coding, which degrades the prediction accuracy. To tackle this problem, we adopt the online training scheme in the second stage classification to refine samples predicted as the FULL RDO by the first stage classification. Previous encoded INTRA CU samples are statistically used in training models for the online classifiers [24]. Since the sample collection and training will introduce the complexity overhead in the encoding process, the number of training samples of each DL is set to be smaller than 1000 and the number of training frames is within N_{Train}, whose aim is to maintain the prediction accuracy while minimizing the complexity overhead. Here, N_{Train} is set as 2 to 4 depends on the resolution of test sequences.

Similar to the ET algorithm in Section IV.A, the weighted factors (W_{S}/W_{NS}) for the classifiers (four classifiers in total and one for each DL) in the second stage classification shall also be determined. Five sequences including *BlowingBubbles*, *PartyScene*, *Vidyo3*, *BasketballDrive* and *Traffic* were tested. *QP* was set as 22, 27, 32 or 37, and the first stage offline trained classifier was activated with the pre-determined optimal parameter sets. Eleven sets of the weighted factors (W_{S},W_{NS}), including (5,1), (4,1), (3,1), (2,1) (1,1), (1,2), (1,3), (1,4) and (1,5), were tested. The coding performance are evaluated in terms of the prediction accuracy and ratio of early terminated samples, as shown in Fig. 7.

It is observed from the Fig. 7 that the prediction accuracy of the second stage ET classification reduces as the $\mathrm{Log}_2(W_{\mathrm{S}}/W_{\mathrm{NS}})$ increases, *i.e.*, (W_{S}, W_{NS}) changes from (1,5) to (5,1). High prediction accuracy indicates small RD degradation. The prediction accuracy can be higher than 95% for DL3 and 98% for DL0 to DL2 if we select small $\mathrm{Log}_2(W_{\mathrm{S}}/W_{\mathrm{NS}})$, which maintains sufficient high accuracy. Note that for the CU DL0, the results of *BlowingBubbles* and *PartyScene* are not counted in the final accuracy and ratio results since there is no 64×64 CU in most cases. On the other hand, the ratio of the early terminated samples increases for the four DLs as the parameter $\mathrm{Log}_2(W_{\mathrm{S}}/W_{NS})$ increases. This ratio indicates the complexity reduction of the encoder and more complexity reduction could be achieved as if the ratio becomes larger. Meanwhile, the ratio increases slowly for DL0 but fast for DL3. Basically, the trends of the ET ratio are in inverse proportion to the prediction accuracies. Here, the optimization objective of this stage classification is to minimize the coding complexity subject to negligible RD degradation. In other words, it maximizes the ET ratio subject to sufficient high prediction accuracy, *e.g.*, 97%. The parameter (W_{S}, W_{NS}) from DL0 to DL3 are set as (1:2), (1:2.25), (1:2.5) and (1:4.0), respectively, so as to give high priority to prediction accuracy.

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology

V. Experimental Results and Analyses

To testify the coding effectiveness of the proposed CU size decision approaches, we implemented them and reference algorithms on the latest reference HEVC platform HM16.7 [31]. Let "Proposed_C1" and "Proposed_ALL" denote only the first stage classification is activated and the two stages of classifications are both activated, respectively, for the proposed algorithm. "ShenCSVT14" [14], "MinCSVT15" [15] and "ZhangTII15" [16] are the fast INTRA CU decision algorithms implemented in HM16.7 for comparison. The sizes of largest CU and smallest CU are 64×64 and 8×8, respectively, which means the maximum CU depth is 4. The minimum and maximum RQT transform sizes are 4 and 32, respectively. Four *QP*s, 22, 27, 32 and 37, were used. All INTRA main configuration was adopted and all the frames were encoded with INTRA frames. The rest coding parameter settings follow the CTC [32]. Twenty-six standard video sequences with various resolutions from 416×240 to 2560×1920 and contents were encoded. All the video coding experiments were performed on HPZ840 workstation with Intel Xeon CPU E5-265v3@2.30 GHz and 2.3GHz, 64GB, Windows 7 64-bit professional operating system. Bjonteggard Delta Peak-Signal-to-Noise Ratio (BDPSNR) and BDBR [33] are used to evaluate the RD performance of the proposed and reference schemes while compared with the original HM16.7. In addition, time saving ratio (ΔT) is used to measure the complexity reduction of the tested schemes, and it is defined as

$$\Delta T = \frac{1}{4}\sum_{i=1}^{4}\frac{T_{HM}\left(QP_i\right)-T_{\varphi}\left(QP_i\right)}{T_{HM}\left(QP_i\right)}\times 100\%\,, \qquad (13)$$

where $T_{HM}(QP_i)$ and $T_{\varphi}(QP_i)$ is the coding time of using the original HM16.7 and the scheme φ with QP_i, $\varphi\in$ {MinCSVT15, ShenCSVT14, ZhangTII15, Proposed_C1, Proposed_ALL}.

At the first stage classification, the first frame of five video sequences, including *BQSqure*, *BaskeballDirll*, *FourPeople*, *ParkScene* and *Traffic* were encoded with *QP*s from 22 to 37 by the original HM16.7, in which the best CU size (the ground truth) and feature values are collected as the training set. As the target allowable bit rate increase ΔR_{T} is set as 1% in this paper, the parameters (W_{S},W_{NS}) of training the offline classifiers are set as (1:2.440), (2.161:1), (4.464:1), and (7.497:1) for the ES scheme from DL0 to DL3. Meanwhile, (W_{S},W_{NS}) are set as (1:4.881), (1:3.683), (1:1.587) and (1:1.193), respectively, for the ET scheme from DL0 to DL3, as illustrated in Table IV. At the predicting stage, the trained models are loaded to encode the 26 test sequences. For the second stage classification, N_{train} is set as 2 for large resolution sequences, such as 1280×720, 1920×1280 and 2560×1920, 3 and 4 for small resolutions sequences, *i.e.*, 832×480 and 416×240, which intends to collect more samples for small resolution sequences. $N_{train}+N_{pred}$=200 means the next 196 to 198 frames are used as predicting and this training/predicting process is repeated every 200 frames. Additionally, for training the online classifier, the parameter (W_{S}, W_{NS}) from DL0 to DL3 are set as (1:2), (1:2.25), (1:2.5)

Table VI

RD and Complexity Comparison between the Proposed Algorithm and the Benchmarks.

Test Sequences	ShenCSVT14 vs HM			MinCSVT15 vs HM			ZhangTII15 vs HM		
	BDBR [%]	BDPSNR [dB]	ΔT [%]	BDBR [%]	BDPSNR [dB]	ΔT [%]	BDBR [%]	BDPSNR [dB]	ΔT [%]
BQMall	1.92	-0.10	-35.06	1.55	-0.08	-25.72	1.05	-0.06	-36.10
BQSquare*	0.25	-0.02	-26.64	0.38	-0.03	-34.41	1.32	-0.11	-23.38
BQTerrace	1.47	-0.07	-40.64	1.79	-0.09	-30.58	1.03	-0.05	-46.96
BasketballDrillText	1.59	-0.08	-32.79	1.54	-0.08	-29.55	0.29	-0.02	-31.19
BasketballDrill*	1.33	-0.06	-32.10	1.84	-0.09	-26.27	0.36	-0.02	-31.07
BasketballDrive	2.44	-0.07	-52.60	2.45	-0.07	-23.20	0.67	-0.02	-48.07
BasketballPass	1.15	-0.06	-35.20	1.01	-0.06	-21.60	0.91	-0.05	-41.21
BlowingBubbles	0.50	-0.03	-20.71	0.75	-0.05	-19.65	0.42	-0.02	-21.45
Cactus	2.29	-0.08	-40.14	1.72	-0.06	-26.23	1.32	-0.04	-42.05
Flowervase	2.48	-0.15	-59.90	0.94	-0.06	-38.80	1.16	-0.07	-53.87
FourPeople*	2.82	-0.15	-48.04	2.52	-0.14	-35.61	1.09	-0.06	-43.73
Johnny	4.43	-0.18	-57.04	3.35	-0.13	-44.81	1.17	-0.05	-55.94
Keiba	1.75	-0.08	-43.19	2.29	-0.11	-30.43	0.96	-0.05	-45.92
Kimono	1.45	-0.05	-36.34	5.05	-0.17	-22.96	1.00	-0.04	-55.26
KristenAndSara	4.80	-0.23	-57.82	3.30	-0.16	-43.02	1.15	-0.06	-54.78
NebutaFestival	1.17	-0.07	-9.90	6.61	-0.40	-17.18	0.41	-0.03	-59.69
ParkScene*	1.72	-0.07	-37.03	1.26	-0.05	-20.91	0.73	-0.03	-37.79
PartyScene	0.40	-0.03	-28.67	0.57	-0.04	-24.10	0.91	-0.06	-30.77
PeopleOnStreet	1.66	-0.09	-36.43	3.75	-0.20	-31.01	0.52	-0.03	-36.23
RaceHorses	0.75	-0.04	-20.34	1.12	-0.06	-20.82	1.86	-0.10	-28.50
SteamLocomotiveTrain	1.69	-0.06	-47.24	3.69	-0.14	-26.28	0.97	-0.04	-60.54
Tennis	3.34	-0.10	-59.49	2.87	-0.08	-31.05	1.22	-0.04	-58.57
Traffic*	1.86	-0.09	-39.39	2.50	-0.12	-24.49	0.72	-0.04	-39.75
Vidyo1	3.39	-0.16	-56.92	3.07	-0.15	-35.63	1.42	-0.08	-51.40
Vidyo3	7.06	-0.36	-59.22	3.05	-0.16	-39.30	0.86	-0.05	-49.80
Vidyo4	4.52	-0.21	-54.12	2.39	-0.11	-35.91	1.10	-0.05	-46.67
Average exclude *	**2.39**	**-0.11**	**-42.08**	**2.52**	**-0.12**	**-29.42**	**0.97**	**-0.05**	**-45.47**
Average overall	**2.24**	**-0.10**	**-41.04**	**2.36**	**-0.11**	**-29.21**	**0.95**	**-0.05**	**-43.49**

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology

TABLE VI
RD AND COMPLEXITY COMPARISON BETWEEN THE PROPOSED ALGORITHM AND THE BENCHMARKS.(CONT.)

Test Sequences	Proposed_C1 vs HM			Proposed_ALL vs HM		
	BDBR [%]	BDPSNR [dB]	ΔT [%]	BDBR [%]	BDPSNR [dB]	ΔT [%]
BQMall	0.98	-0.05	-37.59	1.07	-0.06	-38.38
BQSquare*	0.33	-0.03	-35.96	0.50	-0.05	-36.29
BQTerrace	0.88	-0.05	-42.29	1.05	-0.05	-51.03
BasketballDrillText	0.62	-0.04	-39.11	0.82	-0.05	-39.74
BasketballDrill*	0.76	-0.04	-38.80	0.99	-0.05	-39.74
BasketballDrive	1.31	-0.04	-50.85	1.87	-0.06	-61.09
BasketballPass	0.53	-0.03	-41.73	1.34	-0.08	-45.99
BlowingBubbles	0.15	-0.01	-27.15	0.48	-0.03	-27.95
Cactus	0.95	-0.03	-43.83	1.02	-0.04	-45.50
Flowervase	1.53	-0.09	-56.04	2.22	-0.14	-60.45
FourPeople*	1.53	-0.09	-50.14	1.70	-0.10	-51.76
Johnny	2.16	-0.09	-63.39	3.01	-0.12	-67.99
Keiba	1.00	-0.05	-50.35	1.45	-0.07	-52.23
Kimono	1.08	-0.04	-63.04	3.72	-0.13	-80.53
KristenAndSara	1.83	-0.09	-61.58	2.39	-0.12	-63.56
NebutaFestival	1.47	-0.09	-34.46	1.76	-0.11	-70.46
ParkScene*	0.59	-0.02	-39.88	0.67	-0.03	-40.01
PartyScene	0.16	-0.01	-28.91	0.24	-0.02	-28.82
PeopleOnStreet	1.12	-0.06	-44.37	1.20	-0.06	-44.81
RaceHorses	0.67	-0.04	-33.40	1.18	-0.07	-40.11
SteamLocomotiveTrain	1.17	-0.04	-66.29	1.45	-0.05	-75.93
Tennis	1.67	-0.05	-63.41	2.28	-0.07	-70.81
Traffic*	0.90	-0.05	-45.32	0.98	-0.05	-45.69
Vidyo1	1.95	-0.10	-59.74	2.54	-0.12	-62.01
Vidyo3	2.22	-0.12	-56.02	3.15	-0.17	-64.36
Vidyo4	1.42	-0.07	-56.15	1.89	-0.09	-59.18
Average exclude *	**1.19**	**-0.06**	**-48.56**	**1.72**	**-0.08**	**-54.81**
Average overall	**1.12**	**-0.05**	**-47.30**	**1.58**	**-0.08**	**-52.48**

Note * represents that a few frames of these sequences have been used as training set for offline learning the 1st stage classifiers.

and (1:4.0) respectively, as presented in Section IV.B. For both the offline and online classifier learning, these parameters (W_S, W_{NS}) were determined offline and only once. Therefore, there is no complexity overhead of the encoder for the optimal parameter determination.

Table VI presents the RD and complexity reduction among ShenCSVT14, MinCSVT15, ZhangTII15, the proposed algorithm and the original HM algorithm. It is observed that the ShenCSVT14 reduces complexity from 26.64% to 59.49%, 41.04% on average. Meanwhile, the average BDBR and BDPSNR are 2.24% and -0.10dB, respectively. For sequences, such as *KristenAndSara*, *Johnny*, *Vidyo4* and *Vidyo3*, the BDBR increases from 4.43% to 7.06%, which is large. MinCSVT15 scheme reduces the coding time from 17.18% to 44.81%, 29.21% on average. Meanwhile, the average BDBR increase and BDPSNR degradation are 2.36% and 0.11 dB, respectively. For sequences such as *NebutaFestival* and *Kimono*, the complexity reduces 17.18% and 22.96%, while the BDBRs increases 6.61% and 5.05%, respectively. This method is not so efficient for these sequences mainly because the hard threshold is not effective for all sequences and various frames. In addition, we can observe that ZhangTII15 reduces complexity from 23.36% to 55.94%, and 43.49% on average, while the average BDBR and BDPSNR are 0.95% and -0.05 dB, respectively.

As for the proposed algorithms, it is observed that the Proposed_C1 achieves 27.15% to 66.29%, 47.30% complexity reduction on average when compared with the original HM. Meanwhile, the BDBR and BDPSNR are 1.16% and -0.05 dB on average. While excluding the five training sequences, the average complexity reduction, BDBR and DBPSNR are 48.56%, 1.19% and -0.06 dB, respectively. As for the Proposed_ALL in which the two stages of classification are jointly used, it can achieve 36.29% to 80.53% complexity reduction on average. Meanwhile, the BDBR and BDPSNR are 1.58% and -0.08 dB on average, respectively. While excluding the five training sequences, the average complexity reduction, BDBR and DBPSNR are 54.81%, 1.72% and -0.08dB. Compared with the Proposed_C1, the Proposed_ALL can achieve 52.48% - 47.30%=5.18% more complexity reduction, which means the proposed algorithm becomes more effective owing to the online learning. Especially for the large resolution sequences including *Kimono*, *BasketballDrive*, *NebutaFestival*, *SteamLocomotiveTrain* and *Tennis etc.*, the Proposed_ALL can achieve about 10% more complexity reduction. Compared with ShenCSVT14 and MinCSVT15, the Proposed_ALL achieves much more complexity reduction and less RD degradation. As compared with ZhangTII15, the Proposed_C1 has 3.81% more complexity reduction while at cost of 0.17% BDBR increase. Meanwhile, the Proposed_ALL has 52.48% - 43.49% = 8.99% more complexity reduction at cost of 0.03 dB BDPSNR degradation or 0.63% BDBR increase. Basically, the Proposed_C1 and Proposed_ALL have more complexity reduction and are more effective.

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology

TABLE VII
COMPLEXITY OVERHEAD OF THE CLASSIFICATION AT DIFFERENT STAGES AND DLS. [%]

Test Sequences	$C_{OH}(\phi)$ of 1st Stage Offline Classification					$C_{OH}(\phi)$ of 2nd Stage Online Classification
	DL0	DL1	DL2	DL3	Total	
BQMall	2.10	2.51	4.11	6.05	14.77	2.10
BQSquare	1.97	2.53	4.19	6.73	15.42	0.79
BQTerrace	2.32	2.47	4.15	5.80	14.74	2.91
BasketballDrillText	2.22	2.68	4.40	5.74	15.03	1.69
BasketballDrill	2.24	2.69	4.41	5.62	14.95	1.88
BasketballDrive	2.82	2.90	4.27	5.03	15.02	2.21
BasketballPass	1.86	2.67	4.49	5.47	14.49	0.96
BlowingBubbles	1.48	2.19	3.89	6.44	13.99	1.21
Cactus	2.31	2.53	4.05	5.59	14.47	2.58
Flowervase	2.94	3.12	4.11	4.54	14.72	0.50
FourPeople	2.58	2.75	4.68	5.82	15.83	1.73
Johnny	3.50	2.55	4.47	5.04	15.56	0.91
Keiba	2.63	2.60	4.23	6.36	15.82	2.06
Kimono	3.12	3.59	4.94	3.85	15.50	0.66
KristenAndSara	3.39	3.05	4.28	4.82	15.54	0.83
NebutaFestival	1.87	2.02	3.34	5.53	12.75	2.21
ParkScene	2.18	2.37	3.90	5.43	13.88	2.77
PartyScene	1.86	2.12	3.57	6.26	13.81	2.20
PeopleOnStreet	2.34	2.67	4.87	6.60	16.48	2.57
RaceHorses	1.61	2.07	4.05	6.03	13.75	1.13
SteamLocomotiveTrain	3.72	2.79	3.71	3.95	14.17	1.33
Tennis	3.35	3.41	4.82	4.25	15.83	1.40
Traffic	2.39	2.68	4.61	5.64	15.32	2.56
Vidyo1	3.17	3.30	4.93	4.80	16.21	1.08
Vidyo3	2.97	3.18	4.55	5.00	15.71	0.97
Vidyo4	2.94	3.17	5.12	4.92	16.15	1.15
Average	**2.53**	**2.72**	**4.31**	**5.44**	**15.00**	**1.63**

In addition to the coding complexity reduction and coding efficiency, the complexity overhead ($C_{OH}(\phi)$) of classifiers in each stage and DLs are also analyzed, as shown in Table VII. The complexity overhead is defined as

$$C_{OH}\left(\phi\right)=\frac{1}{4}\sum_{i=1}^{4}\frac{T_{\phi}\left(QP_i\right)}{T_{\text{Proposed\_ALL}}\left(QP_i\right)}\times 100\% \qquad (13)$$

where $T_{\text{Proposed\_ALL}}(QP_i)$ is the total time of the Proposed_ALL encoder. $T_{\phi}(QP_i)$ is the time cost of the learning algorithms, including feature extraction, classifier prediction, and training for the online learning. ϕ is the classifications at DL*i* and total of the first stage classification, as well as total of the second stage classification. It is observed that average complexity overhead from DL0 to DL3 are 2.53%, 2.72%, 4.31% and 5.44%, respectively, and 15.00% in total for the first stage classification. The major overhead comes from the frequent call of feature extraction (especially the features from pre-analysis) and SVM prediction. The overhead of the second stage online classification is 1.63% on average, which is much smaller compared with that of the first stage. This lies on three main reasons. Firstly, the number of FULL RDO samples is much smaller, which is about 3.67% to 41.60% on average (see Table V). Secondly, we strictly control the number of online training samples and N_{train} which is 2 for large resolution sequences. It reduces the number of supporting vectors. Thirdly, feature extraction of texture and context information is implemented and counted in the first stage classification. Meanwhile, getting features on coding information of the current CU is negligible.

VI. CONCLUSIONS

In this paper, we propose effective data driven Coding Unit (CU) size decision approaches for HEVC INTRA coding, which consists of two stages of learning based fast INTRA CU decisions at each CU decision layer. At the first stage classification, a three output classifier with offline learning is developed to early terminate the CU size decision or early skip the current CU depth. As for the samples neither early terminated nor early skipped in first stage, binary classification at the second stage, which is online trained from previous coded frames, is proposed to further enhance the CU size decision. Representative features are explored for different decision layers and stages of classifications. Finally, the optimal parameters derived from the training data are achieved to reasonably allocate complexity among different CU layers at given total rate-distortion degradation constraint. Extensive experiments demonstrate that the proposed algorithm outperforms the state-of-the-art schemes in complexity reduction and is effective.

REFERENCES

[1] G.J. Sullivan, J.R. Ohm, W.J. Han, and T. Wiegand, Overview of the High Efficiency Video Coding (HEVC) Standard, *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 22, No. 12, pp.1649-1668, Dec. 2012.

[2] Y. Zhang, S. Kwong, G. Jiang, X. Wang, and M. Yu, Statistical Early Termination Model for Fast Mode Decision and Reference Frame

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology

Selection in Multiview Video Coding, *IEEE Trans. Broadcast.*, vol.58, no.1, pp.10-23, March 2012.

[3] Z Pan, Y. Zhang, and S. Kwong, Efficient Motion and Disparity Estimation Optimization for Low Complexity Multiview Video Coding, *IEEE Trans. Broadcast.*, Vol. 61, No. 2, June 2015, pp.166-176.

[4] Z. Pan, J. Lei, Y. Zhang, X. Sun, and S. Kwong, "Fast Motion Estimation Based on Content Property for Low-Complexity H.265/HEVC Encoder", *IEEE Trans. Broadcast.*, Vol. 62, No. 3, Sept. 2016, pp. 675 - 684.

[5] L. Shen, Z. Zhang and Z. Liu, Adaptive Inter-mode Decision for HEVC Jointly Utilizing Inter-level and Spatio-Temporal Correlations, *IEEE Trans. Circuits Syst. Video Technol.*, Vol.24, No.10, pp.1709-1722, Sep. 2014.

[6] Y. Zhang, S. Kwong, X. Wang, Z. Pan, H. Yuan, and L. Xu, Machine Learning Based Coding Unit Depth Decisions for Flexible Complexity Allocation in High Efficiency Video Coding, *IEEE Trans. Image Process.*, vol.24, no.7, pp.2225-2238, July. 2015.

[7] X. Shen, and L. Yu, CU Splitting Early Termination Based on Weighted SVM, *EURASIP J Image Video Process.*, vol. 2013, article 4, 2013.

[8] L. Zhu, Y. Zhang, N. Li, G. Jiang, S. Kwong, Machine Learning Based Fast H.264/AVC to HEVC Transcoding Exploiting Block Partition Similarity, *J. Vis. Commun. Image R.*, Vol. 38, July 2016, pp. 824–837,

[9] L. Zhu, Y. Zhang, Z. Pan, R. Wang, S. Kwong, and Z. Peng, "Binary and Multi-Class Learning Based Low Complexity Optimization for HEVC Encoding," *IEEE Trans. Broadcast.,* May, 2017, DOI, 10.1109/TBC.2017.2711142.

[10] E. Peixoto, T. Shanableh, and E. Izquierdo, H.264/AVC to HEVC Video Transcoder Based on Dynamic Thresholding and Content Modeling, *IEEE Trans. Circuits Syst. Video Technol.*, Vol.24, No.1, pp.99-112, Jan. 2014.

[11] J. Xiong, H. Li, F. Meng, S. Zhu, Q. Wu, and B. Zeng, MRF-Based Fast HEVC Inter CU Decision With the Variance of Absolute Differences, *IEEE Trans. Multimedia*, Vol. 16, No. 8, pp.2141-2153, Dec. 2014.

[12] G. Correa, P.A. Assuncao, L.V. Agostini, and L.A. Cruz, Fast HEVC Encoding Decisions Using Data Mining, *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 25, No. 4, Apr. 2015, pp.660-673.

[13] J. Lainema, F. Bossen, W.J. Han, J. Min, and K. Ugur, Intra Coding of the HEVC Standard, *IEEE Trans. Circuits Syst. Video Technol.*, Vol.22, No.12, Dec. 2012, pp. 1792-1801.

[14] L. Shen, Z. Zhang, and Z. Liu, Effective CU Size Decision for HEVC Intracoding, *IEEE Trans. Image Process.*, Vol. 23, No. 10, Oct. 2014, pp.4232-4241.

[15] B. Min and R. C. C. Cheung, A Fast CU Size Decision Algorithm for the HEVC Intra Encoder, *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 25, No. 5, May 2015, pp. 892-896.

[16] Y. Zhang, S. Kwong, G. Zhang, Z. Pan, Y. Hui, and G. Jiang, Low Complexity HEVC INTRA Coding for High Quality Mobile Video Communication, *IEEE Trans. Industrial Informatics*, Vol.11, No.6, Dec. 2015, pp.1492-1504.

[17] X. Zuo, L. Yu, Fast Mode Decision Method for All Intra Spatial Scalability in SHVC, *IEEE Conf' Visual Commun. Image Process. (VCIP)*, Valletta, Malta, Dec. 2014, pp. 394-397.

[18] C. S. Park, "Efficient Intra-mode Decision Algorithm Skipping Unnecessary Depth-modelling Modes in 3D-HEVC", *Electron. Lett.*, Vol. 51, No.10, 2015, pp. 756 - 758

[19] Q. Zhang, X. Huang, X. Wang and W. Zhang, A Fast Intra Mode Decision Algorithm for HEVC Using Sobel Operator in Edge Detection, *Int'l J. Multimedia Ubiquitous Engineering*, Vol.10, No.9, 2015, pp.81-90

[20] L. L. Wang, and W. C. Siu, Novel Adaptive Algorithm for Intra Prediction with Compromised Modes Skipping and Signaling Processes in HEVC, *IEEE Trans. Circuits Syst. Video Technol.*, Vol.23, No.10, pp.1686-1694, Oct. 2013.

[21] H. Zhang, and Z. Ma, Fast Intra Mode Decision for High Efficiency Video Coding (HEVC), *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 24, No. 4, Apr. 2014, pp.660-668.

[22] K. Lim, J. Lee, S. Kim, and S. Lee, Fast PU Skip and Split Termination Algorithm for HEVC Intra Prediction, *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 25, No. 8, Aug. 2015, pp. 1335-1346

[23] N. Hu, and E. H. Yang, Fast Mode Selection for HEVC Intra Frame Coding with Entropy Coding Refinement Based on Transparent Composite Model, *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 25, No. 9, Sept. 2015, pp.1521 - 1532.

[24] J. Chen and L. Yu, Effective HEVC Intra Coding Unit Size Decision based on Online Progressive Bayesian Classification, *IEEE Int'l Conf. Multimedia Expo (ICME)*, Seattle, WA, USA, 2016, pp. 1-6.

[25] B. Du, W. C. Siu and X. Yang, Fast CU Partition Strategy for HEVC Intra-frame Coding Using Learning Approach via Random Forests, *Asia-Pacific Signal Info. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Hong Kong, 2015, pp. 1085-1090.

[26] D. R. Coll, V. Adzic, G. F. Escribano, H. Kalva, J.L. Mart ńez, P. Cuenca, Fast Partitioning Algorithm for HEVC Intra Frame Coding using Machine Learning, *IEEE Int'l Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 4112-4116.

[27] F. Duanmu, Z. Ma and Y. Wang, Fast Mode and Partition Decision Using Machine Learning for Intra-Frame Coding in HEVC Screen Content Coding Extension, *IEEE J. Emerg. Select. Topics Circuits Syst.*, 2016, Volume: PP, Issue: 99, pp.1 - 15, DOI: 10.1109/JETCAS.2016.2597698 (in press)

[28] T. Zhang, M.T. Sun, D. Zhao, and W. Gao, Fast Intra Mode and CU Size Decision for HEVC, *IEEE Trans. Circuits Syst. Video Technol.*, 2016, DOI: 10.1109/TCSVT.2016.2556518(in press)

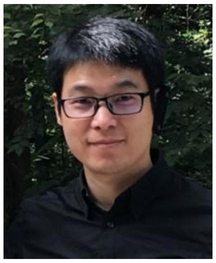
[29] Z. Liu, X. Yu, Y. Gao, S. Chen, X. Ji, and D. Wang, CU Partition Mode Decision for HEVC Hardwired Intra Encoder Using Convolution Neural Network, *IEEE Trans. Image Process.*, 2016, DOI:10.1109/TIP.2016. 2601264. (in press)

[30] C.-C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machines, *ACM Trans. Intelligent Syst. Technol.*, 2:27:1--27:27, 2011.

[31] HEVC Reference Software HM16.7. (2015). Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.7/, Oc. 2015.

[32] F. Bossen, Common Test Conditions and Software Reference Configurations, JCTVC of ISO/IEC and ITU-T, JCTVC-J1100, Stockholm, SE, Jul. 2012.

[33] G. Bj øntegaard, Calculation of Average PSNR Differences between RD-curves, ITU-T SG16 Q.6, VCEG-M33, Austin, TX, Apr. 2001.



Yun Zhang (M'12-SM'16) received the B.S. and M.S. degrees in electrical engineering from Ningbo University, Ningbo, China, in 2004 and 2007, respectively, and received the Ph.D. degree in computer science from Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2010. From 2009 to 2014, he was a Post-doc Researcher with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong. In 2010, he joined in the Shenzhen Institutes of Advanced Technology (SIAT), CAS, as an Assistant Professor. Since 2012, he serves as Associate Professor. His research interests are 3D video, high efficiency video coding and perceptual video processing.

Zhaoqing Pan (M'14) received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2014. In 2013, he was a Visiting Scholar with the Department of Electrical Engineering, University of Washington, Seattle, WA, USA, for six months. He is currently a Professor with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China. His research interests include video compression and 3-D video processing.

Na Li received the B.S. degree in computer science and technology from Hunan University, Changsha, China, in 2009, and the Ph.D. degree in computer application and technology from the Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, in 2014. From 2014 to 2016, she was a Post-Doctoral Researcher with the Center for High Performance Computing, Shenzhen Institute of Advanced Technology (SIAT), CAS, Shenzhen, China. In 2016, she became an Assistant Professor with SIAT, CAS. Her research interests are machine learning, crowded scene analysis, video analysis and video processing.

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2747659, IEEE Transactions on Circuits and Systems for Video Technology

Xu Wang (M'15) received the B.S. degree from South China Normal University, Guangzhou, China, in 2007, and M.S. degree from Ningbo University, Ningbo, China in 2010. He received his Ph.D. degree from the Department of Computer Science, City University of Hong Kong, Hong Kong in 2014. In 2015, he joined the College of Computer Science and Software Engineering, Shenzhen University as an Assistant Professor. His research interests are video coding and stereoscopic image/video quality assessment.

Gangyi Jiang (M'10) received his M.S. degree in electronics engineering from Hangzhou University in 1992, and received his Ph.D. degree in electronics engineering from Ajou University, Korea, in 2000. He is now a professor in Faculty of Information Science and Engineering, Ningbo University, China. His research interests mainly include video compression, multi-view video coding, etc. He has published over 100 technical articles in refereed journals and proceedings in these fields.

Sam Kwong (M'93–SM'04-F'13) received the B.S. and M.S. degrees in electrical engineering from the State University of New York at Buffalo in 1983, the University of Waterloo, Waterloo, ON, Canada, in 1985, and the Ph.D. degree from the University of Hagen, Germany, in 1996. From 1985 to 1987, he was a Diagnostic Engineer with Control Data Canada. He joined Bell Northern Research Canada as a Member of Scientific Staff. In 1990, he became a Lecturer in the Department of Electronic Engineering, City University of Hong Kong, where he is currently a Professor in the Department of Computer Science. His research interests are video and image coding and evolutionary algorithms.

1051-8215 (c) 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.