A LOW-PARAMETRIC MODEL FOR BIT-RATE ESTIMATION OF VVC RESIDUAL CODING

Fabian Brand, Christian Herglotz, André Kaup

Multimedia Communications and Signal Processing, Friedrich-Alexander-Universität Erlangen-Nürnberg

ABSTRACT

There are many tasks within video compression which require fast bit rate estimation. As an example, rate-control algorithms are only feasible because it is possible to estimate the required bit rate without needing to encode the entire block. With residual coding technology becoming more and more sophisticated, the corresponding bit rate models require more advanced features. In this work, we propose a set of four features together with a linear model, which is able to estimate the rate of arbitrary residual blocks which were compressed using the VVC standard. Our method outperforms other methods which were used for the same task both in terms of mean absolute error and mean relative error. Our model deviates by less than 4 bit on average over a large dataset of natural images.

***Index Terms*—** Versatile Video Coding, Rate Model, Rate Control

1. INTRODUCTION

Bit rate estimation is a common problem in image and video compression. There are multiple tasks in a video coder where knowledge of the resulting bit rate is beneficial. This includes for example rate-control [1], where the encoder has to match a specified rate at the output. Here, the encoder has to know the resulting rate of a coding decision in order to accurately choose the coding parameters for optimal rate-distortion behavior, a process which is known as rate-distortion-optimization [2, 3]. One possibility to obtain the rate needed for a certain decision is to perform the compression until the end and measure the rate directly. While this is certainly the most accurate method, following the coder to the end is often time-consuming, when performed too often. In such cases, a robust, simple and accurate rate model is useful to estimate the required rate.

One example where the rate has to be estimated multiple times is rate-distortion optimized quantization (RDOQ) [4, 5, 6]. Here the quantized values themselves are subject to rate-distortion optimization. For each tested configuration, both rate and distortion have to be evaluated. Another application targets video formats with irrelevant content. For example, some projection formats in 360° video coding map the content to a 2D plane, where parts of the sequence are irrelevant for the reconstruction of the 360° content [7]. Similar regions can be found in videos generated for video-based point cloud compression [8]. In such sequences, several blocks may contain both relevant as well as irrelevant pixels, where the exact pixel values in the irrelevant region are of no importance. Depending on the number of irrelevant pixels, this can lead to a manifold of optimal representations in the DCT domain, where the best representation must be determined for optimal rate savings. Using a simple rate model as proposed in this paper instead of performing the arithmetic encoding chain for all DCT representations will decrease the encoding time significantly.

Coding of frequency coefficients has become more sophisticated with ongoing standard development. While JPEG only uses run-level coding to encode the values [9], VVC uses multiple coding passes and can therefore exploit the sparsity of images in the frequency domain, in particular in high frequency areas. Additionally, VVC uses context adaptive binary arithmetic coding (CABAC) [10], which is able to take the previously coded signals into account to improve compression efficiency. So in the end, the required rate for one residual block follows more complicated relationships and is not even deterministic, since different contexts in CABAC can yield different rates for the same residual signal.

In this work, we limit ourselves to models which estimate the rate needed to transmit a transformed and quantized residual block in VVC. Our goal is to design a rate model, which only uses the quantized frequency coefficients and which is valid for all block-sizes. In the following, we first introduce our model and the features on which it is based. We then compare our model against other methods performing a similar task and perform ablation studies demonstrating the impact of the individual features. In the end, we show the versatility and robustness of our model in various experiments.

2. RELATED WORK

Rate models were often examined in the context of rate-control. This field was dominated for a long time by ρ-domain models [11, 12]. Here, a linear relationship between the number of non-zero frequency coefficients and the rate is assumed. In these models, the proportionality factor is often not constant for all contents but rather determined dynamically based on various image and video properties, like texture or previously coded pictures [12].

Another class of rate-control models are the so-called λ-domain models [13, 14]. These models assume a usually ex-

ponential relationship between the rate and the Lagrangian factor λ of the rate-distortion optimization. As for the ρ-domain model, the parameters of the model are often estimated using various image properties.

Recently, learning-based methods for the use in video coding have been proposed. These methods often tackle prediction problems, such as [15] which proposes a deep-learning-based intra-prediction method for use in VVC. Also, VVC includes Matrix Intra Prediction (MIP) [16], which includes a trained matrix. These methods generate prediction signals which are optimal if the residual can be compressed well. These methods therefore require a loss function which accurately estimates the rate of a residual block. Here, it is desirable to have fixed model parameters unlike the ρ- and λ-domain models, that do not depend on other image characteristics. One such model was proposed in [17] which was also used to train MIP.

3. RATE ESTIMATION MODEL

The rate estimation model we propose in this work consists of four features and a bias, which form a linear model with five parameters. We chose a linear model due to its low complexity and simplicity in training. With only five parameters, we expect the model to be robust against overfitting and to generalize well. The features are hand-crafted and inspired by the process of residual coding in VTM. The features are all computed on sub-block level, meaning, each block is divided into 4×4 sub-blocks before feature computation. The feature of the block $\mathcal{B}$ itself is then the sum of the features of all sub-blocks. This design concept grants us the possibility to find a model which is valid for all block-sizes. In the following, the set of all coefficients c in a sub-block is denoted as $\mathcal{S}$.

As first feature we choose to use the number of non-zero coefficients after quantization. This feature is similar to the parameter of the ρ-domain model. The motivation of the feature is that in the beginning of the compression scheme, non-zero coefficients are signaled and all subsequent bits are only required for there so-called significant coefficients. This has been proven a good feature in the past, in particular in the ρ-domain model. In the following, this feature is denoted as S. We can express this feature as

$$S = \sum_{\mathcal{S} \in \mathcal{B}} \sum_{c \in \mathcal{S}} \begin{cases} 0 & \text{if } c \neq 0 \\ 1 & \text{else} \end{cases}. \tag{1}$$

As second feature we use the sum of the binary logarithm of all significant coefficients. This feature estimates how much rate has to be spent to transmit each value. We denote this feature as L with:

$$L = \sum_{\mathcal{S} \in \mathcal{B}} \sum_{c \in \mathcal{S}} \max(0, \log_2 |c|). \tag{2}$$

The next two features describe the positions and distributions of the coefficients on sub-block-level. For each sub-block, we perform a zig-zag scan and look at the last coefficient which is greater than zero. We then use the position of that coefficient as sub-block feature and sum up all the positions for the block feature. We denote this third feature as Z.

The fourth feature measures the percentage of coefficients in the sub-block which are strictly greater than 1 and computes the binary entropy function of the percentage:

$$E = \sum_{\mathcal{S} \in \mathcal{B}} H_2 \left(\frac{C_1(\mathcal{S})}{16} \right), \tag{3}$$

where $H_2(p) = -p \log_2(p) - (1-p) \log_2(1-p)$ is the binary entropy function and $C_n(\mathcal{S})$ is a function counting all coefficients with values greater than n in the sub-block $\mathcal{S}$. In practice, this can easily be realized with a lookup table.

Altogether, we use the four features in a linear model for rate estimation:

$$R_{\text{est}} = a \cdot S + b \cdot L + c \cdot Z + d \cdot E + e, \tag{4}$$

with trainable parameters a, b, c, d, and e, the latter of which is a global offset.

4. EXPERIMENTS

4.1. Setup

To evaluate our model, we compare it with two different models which are both estimating the rate based on transformed blocks. First, we use the ρ-domain model from [11]. The ρ-domain model assumes a linear relationship between the percentage of non-zero quantized coefficients $(1-\rho)$ and the rate. Since this model was proposed for usage in MPEG-4, where all transform blocks were of size 4×4 and therefore constant, the model does not take the size of the block into account. We instead use the absolute number of non-zero coefficients as a feature.

Also, we use the model which was suggested in [17] as loss function to train intra prediction networks for MIP [16]. Since this model was initially proposed as a loss function for a neural network, a linear relationship to the actual rate was sufficient. In our case, however, we require an exact estimate. We therefore extended the model by two parameters, such that the rate for one block is now estimated by

$$R_{\text{est}} = \sum_{(m,n)} \alpha \left| c_{m,n} \right| + \beta g \left(\gamma \left| c_{m,n} \right| + \delta \right) + \varepsilon, \tag{5}$$

with the logistic function $g(\cdot)$. We therefore call this model the logistic model. At this point we want to note that this model, was designed as differentiable model and therefore does not use features based on thresholding, counting, or positions.

For our experiments we encoded 30 picture from the DIV2K dataset [18] using VTM 10.0 [19] and QPs of 22,

		ρ domain [11]				Logistic [17]				Sub-Block Model (Ours)			
QP_t	QP_e	P	MAE	MRE	time	P	MAE	MRE	time	P	MAE	MRE	time
22	22	0.9752	12.3	17.1%	0.003	0.9959	6.91	12.7%	0.002	0.9978	4.78	9.1%	0.009
27	27	0.9845	8.0	17.2%	0.003	0.9936	6.10	15.2%	0.003	0.9970	4.02	10.1%	0.009
32	32	0.9898	5.96	17.5%	0.003	0.9941	4.12	15.7%	0.003	0.9970	3.45	12.7%	0.009
37	37	0.9874	4.55	17.0%	0.003	0.9902	5.33	16.5%	0.004	0.9954	2.98	14.3%	0.009
22	27	0.9830	8.42	17.7%	-	0.9947	5.48	14.6%	-	0.9967	4.07	10.9%	-
	32	0.9905	6.70	18.8%	-	0.9928	4.01	16.1%	-	0.9967	3.52	12.5%	-
37	27	0.9849	8.9	18.1%	-	0.9898	7.76	16.1%	-	0.9968	4.60	11.6%	-
	32	0.9904	6.01	17.9%	-	0.9905	5.84	16.1%	-	0.9969	3.57	12.3%	-

Table 1. Results of the rate estimation experiments. QP_t denotes the QP which was used for training, QP_e denotes the QP which was used for evaluation. The time is given in ms per block.

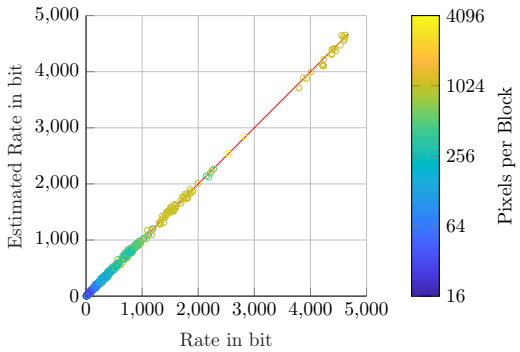


Fig. 1. Estimation curve for our proposed rate model.

27, 32 and 37. This yields between 450,000 and 1,100,000 blocks per QP. We then perform a 5-fold cross validation and average the metrics. We train each model to minimize the mean squared error (MSE). For the ρ-domain model and our proposed method, we can use the pseudo-inverse for estimation, while the non-linear logistic model is trained using gradient descent.

To evaluate the results, we use the Pearson correlation coefficient

$$P = \frac{\sigma_{R,R_{\text{est}}}}{\sigma_R \sigma_{R_{\text{est}}}}, \tag{6}$$

where $\sigma_{R,R_{\text{est}}}$ denotes the covariance of R and R_{est}. We furthermore evaluate the mean absolute error MAE

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| R_i - R_{\text{est},i} \right| \tag{7}$$

and the mean relative error MRE.

$$\text{MRE} = \frac{1}{N} \sum_{i=1}^{N} \frac{\left| R_i - R_{\text{est},i} \right|}{R_i}, \tag{8}$$

where R_i and $R_{\text{est},i}$ denote the i^{th} sample in the test set and N is the total number of tested blocks.

4.2. Rate Estimation

In a first experiment, we compare the three models for each QP separately. In the 5-fold cross validation, both training set and test set were compressed with the same QP. We show the results evaluated on the test sets in Tab. 1.

Feature Set	P	MAE	MRE
S	0.9791	12.34	17.2%
L	0.8995	33.03	48.36%
Z	0.9242	20.69	27.1%
E	0.9361	25.02	45.8%
$S+L$	0.9976	5.70	11.8%
$S+Z$	0.9820	12.38	18.8%
$S+E$	0.9844	10.32	16.7%
$L+Z$	0.9934	7.67	14.9%
$L+E$	0.95	19.92	36.8%
$Z+E$	0.9710	15.11	26.7%
$S+L+Z$	0.9977	5.3	11.2%
$S+L+E$	0.9975	5.34	10.9%
$S+Z+E$	0.9878	10.40	16.6%
$Z+L+E$	0.9948	6.76	13.6%
$S+Z+L+E$	0.9978	4.93	10.5%

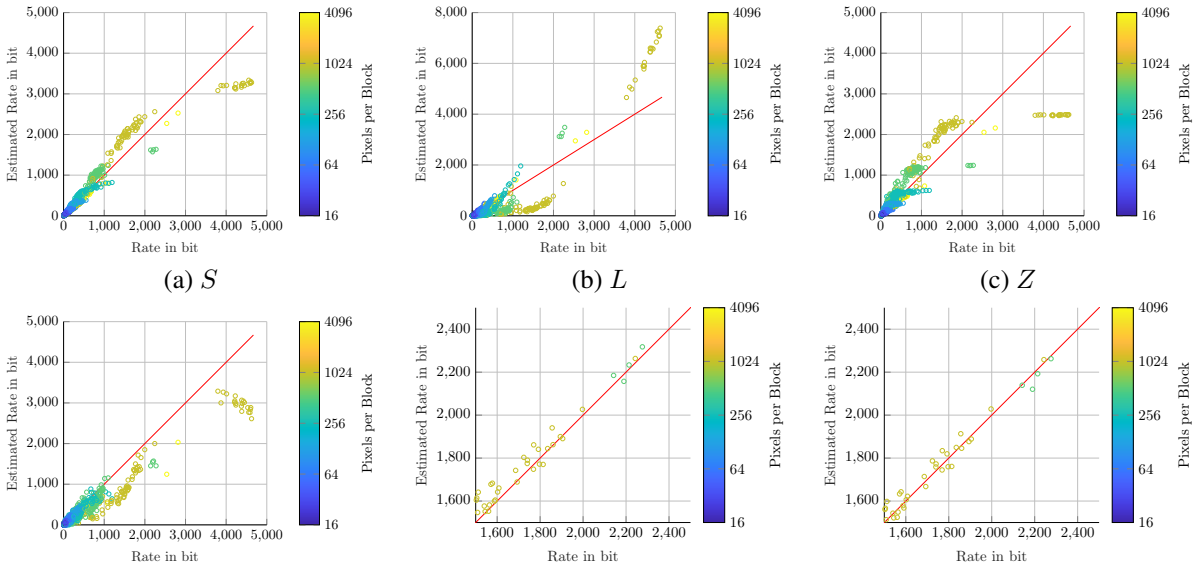
Table 2. Results of the ablation studies.

From this result we see that the ρ-domain model, even though it is a very simple model, still performs relatively well. This indicates that the number of non-zero coefficients is a good indicator and also works well if the parameters are trained on general images and not on specific blocks.

The sigmoid model, which is a more complex model and takes the magnitude of the coefficients into account, performs better with a mean relative error of 12.1% for QP 22 and 19.5% for QP 37. The rise of the MRE with the QP indicates that the model does not perform well for small rates, since the rate average rate decreases with QP.

Our proposed sub-block model performs better than the other two models in all metrics. Throughout all QPs, we achieve a correlation coefficient of $R > 0.995$ and the relative error—even though it is also increasing with the QP—is between 2 and 5 percentage points below the MRE of the sigmoid model. Fig. 1 shows the measured bit rate against the estimated rate. The points are color coded according to the total number of pixels of that block. We see that our model is able to produce accurate results over all rates. The red line indicates the case of perfect estimation.

To compare the complexity of our methods, we conducted



(a) S (b) L (c) Z

(d) E (e) $S + L + Z$ (zoomed) $S + Z + L + E$ (zoomed)

Fig. 2. Estimation curves for different feature combinations.

time measurements in our MATLAB implementation. Here, we see that our method takes about 3 times longer than the others. Note, that the runtime may vary in different implementations. The measurements show that our method is in a similar range than the others with several μs per block.

To demonstrate the versatility of our model, we additionally perform a cross QP test. For this test, we train the models on images which were coded with $\mathrm{QP_t} \in \{22, 37\}$ and evaluate on images which were coded with $\mathrm{QP_e} \in \{27, 32\}$. The results show that all three methods are relatively robust to the QP of the training data. In most cases, we can observe a small degradation when we train with deviating QPs, however, this is always less than 2% in MRE. On the other hand, we also see a few cases where the results are better. Slight deviations are expected since data may randomly produce a better fit. As expected, however, on average the quality degrades with a QP mismatch. In the end, our proposed method still outperforms the other methods in all metrics also in this case.

4.3. Ablation Study

Additionally, we perform ablation studies to demonstrate the descriptive power of each feature. The following experiments were performed with a QP of 22. The results in Tab. 2 clearly show that the individual features are not well suited to estimate the rate. The best individual feature is the number of non-zero coefficients S, which is equivalent to the ρ-domain model. In the table, we can see that especially L and E perform poorly by themselves. In Fig. 2, we show the results of a selection of cases to further illustrate the results.

We can see that S and Z severely underestimate the rate for large rates and that the behavior is non-linear. This is due to the fact that both features do not take the actual values of the coefficients into account. On the other hand, the model using only L leads to an overestimation for large rates. In Tab. 2, we see that $S + L$ and $Z + L$ both give reasonably good results as both effects cancel out and form a better linear model. In Fig. 2 (e) and (f), we see show the effect of the entropy based feature. This feature mainly influence the region for medium bit rates. We can see here that in these ranges, $S + L + Z$ leads to a slight overestimation, which can be fixed by taking the entropy-based feature into account. Note that all ablation study experiments were performed without a bias term. This also shows the importance of being able to add a constant term, as it is responsible for the improvement from 10.5% to 9.1% relative error.

5. CONCLUSION

In this paper, we proposed a novel set of features to estimate the rate required to compress a transformed residual block in VVC. Our model has only five parameters and can be trained easily and stably due to its low complexity and linearity. Other than in known approaches like λ-domain models, we can find fixed parameters to model the rate for all blocks, independent of the content.

In future work, the model accuracy could further be increased by taking the CABAC context into account and including it in the model. Since the context can change the required rate, no model which does not include the context can exceed a certain accuracy. Furthermore, the model can be extended to other video compression standards and the effect of our model in several scenarios, such as RDO, RDOQ or irrelevant region coding, both in compression performance and in runtime can be evaluated. The model we presented in this paper showed superior performance to other models, while remaining simple and linear. Therefore, a speedup and good performance can be expected in various applications.

6. REFERENCES

[1] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G.J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 688–703, July 2003.

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

[3] Xiang Li, N. Oertel, A. Hutter, and A. Kaup, "Laplace distribution based lagrangian rate distortion optimization for hybrid video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 2, pp. 193–205, Feb. 2009.

[4] K. Ramchandran and M. Vetterli, "Rate-distortion optimal fast thresholding with complete JPEG/MPEG decoder compatibility," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 700–704, Sept. 1994.

[5] Jakub Stankowski, Cezary Korzeniewski, Marek Domanski, and Tomasz Grajek, "Rate-distortion optimized quantization in HEVC: Performance limitations," in *Proc. Picture Coding Symposium (PCS)*, May 2015.

[6] Heiko Schwarz, Tung Nguyen, Detlev Marpe, Thomas Wiegand, Marta Karczewicz, Muhammed Coban, and Jie Dong, "Improved quantization and transform coefficient coding for the emerging versatile video coding (VVC) standard," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Sept. 2019.

[7] C. Herglotz, M. Jamali, S. Coulombe, C. Vazquez, and A. Vakili, "Efficient coding of 360° videos exploiting inactive regions in projection formats," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 1104–1108.

[8] L. Li, Z. Li, S. Liu, and H. Li, "Occupancy-map-based rate distortion optimization and partition for video-based point cloud compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 326–338, Jan. 2020.

[9] G.K. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[10] Thomas Wiegand Detlev Marpe, Heiko Schwarz, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 620–636, July 2003.

[11] Zhihai He, Yong Kwan Kim, and S.K. Mitra, "Low-delay rate control for DCT video coding via ρ-domain source modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 8, pp. 928–940, Aug. 2001.

[12] Meng Liu, Yi Guo, Houqiang Li, and Chang Wen Chen, "Low-complexity rate control based on ρ-domain model for scalable video coding," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Sept. 2010.

[13] Bin Li, Houqiang Li, Li Li, and Jinlei Zhang, "λ domain rate control algorithm for high efficiency video coding," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3841–3854, sep 2014.

[14] Victor Sanchez, "Rate control for hevc intra-coding based on piecewise linear approximations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. apr 2018, IEEE.

[15] Fabian Brand, Jurgen Seiler, and Andre Kaup, "Intra-frame coding using a conditional autoencoder," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 354–365, Feb 2021.

[16] Jonathan Pfaff, Björn Stallenberger, Michael Schäfer, Philipp Merkle, Philipp Helle, Tobias Hinz, Heiko Schwarz, Detlev Marpe, and Thomas Wiegand, "CE3: Affine linear weighted intra prediction (CE3-4.1, CE3-4.2) , JVET-N0217-v1," *14th Meeting of the Joint Video Exploration Team (JVET)*, pp. 1–18, Mar 2019.

[17] Philipp Helle, Jonathan Pfaff, Michael Schäfer, Roman Rischke, Heiko Schwarz, Detlev Marpe, and Thomas Wiegand, "Intra picture prediction for video coding with neural networks," in *Proc. Data Compression Conference (DCC)*, Mar 2019, pp. 448–457.

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

[19] Jianle Chen, Y. Ye, and S. Kim, "Algorithm description for versatile video coding and test model 10 (VTM 10), JVET-S2002," *19th Meeting of the Joint Video Exploration Team (JVET)*, pp. 1–67, Jan 2020.