Visual Saliency Based Perceptual Video Coding in HEVC

Henglu Wei[#], Xin Zhou[*], Wei Zhou[#], Chang Yan[#], Zhemin Duan[#] and Nana Shan[#]
[#]School of Electronics and Information
[*]School of Automation
Northwestern Polytechnical University
Xi'an 710072, China
xinzhou@nwpu.edu.cn

***Abstract*—Perceptual video coding has the potential to provide the same visual quality at a lower bit-rate, compared with the traditional objective quality based scheme. Visual saliency represents the probability of human attention over frames, and it is used for allocating coding bits or controlling visual quality. In this paper, a HEVC compliant perceptual video coding scheme is proposed based on visual saliency. At first visual saliency map is attained to indicate the distribution of saliency. Then refined distortion allocating method is performed in CU level with adaptive QP which is adjusted by the average visual saliency. Besides, a fast CU mode decision algorithm suitable for perceptual video coding in HEVC is proposed to accelerate the encoder. In the fast algorithm, the average saliency is used to estimate texture complexity and movements in videos. Experimental results show that up to 22.52% bit-rate and 43.48% encoding time can be saved by our methods with negligible perceptual quality loss.**

Keywords—visual saliency; HEVC; perceptual video coding

I. INTRODUCTION

High Efficiency Video Coding (HEVC) [1] is the state of art video coding standard jointly developed by the ISO/IEC Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG). As a successor to H.264/AVC, HEVC aims to reduce the bit-rate by a half with the same video quality compared to its predecessor. Although HEVC provides higher compression ratio, it also introduces more complexity than H.264/AVC.

For almost all video encoders, including HEVC, Sum of Squared Error (SSE) is used as a distortion metric while Peak to Signal Noise Ratio (PSNR) as a quality metric. However, it is well known that the SSE metric and PSNR metric are not entirely consistent in visual quality with the perception characteristics of the human visual system (HVS). In contrast, perceptual video coding makes the most of HVS, and it has the potential to reduce more bit-rate. Visual Saliency (VS) is one of the widely used HVS models. VS means a region or an object that is somehow standing out than their neighborhoods. The visual saliency of all pixels in a frame makes up a Visual Saliency Map (VSM).

Some perceptual video coding algorithms have been proposed for H.264/AVC and HEVC. Just Noticeable Distortion (JND) model is used in [2] and [3] to tolerate more distortion. In [4] and [5], VS is used as a metric to allocate visual quality by adjusting Quantization Parameter (QP). VS detection model is combined with a visual sensitivity model in [6] to adjust distortion threshold. There are still two aspects in VS based perceptual video coding needed to be refined. First, the relationship between the degree of VS and distortion is somewhat vague, and more bit-rate can be reduced by a refined quality allocation scheme. Another aspect is that none of them takes the high complexity of HEVC encoder into account. Although some fast mode decision algorithms for HEVC have been developed to accelerate the encoder [7-8], none of them are specially designed for perceptual video coding. What's more, these fast algorithms increase the bit-rate while reduce the encoder complexity. So, fast mode decision algorithms suitable for perceptual coding are in demand.

To overcome the obstacles of the previous work, we propose a refined VS based perceptual video coding scheme for HEVC. In this scheme, perceptual coding is performed in CU level by JND related distortion allocation scheme. Then a fast mode decision algorithm suitable for perceptual coding is designed. The simulation results show that both the encoder complexity and bit-rate of HEVC can be efficiently reduced by the proposed algorithm. The rest of the paper is organized as follows. The proposed distortion allocating scheme and fast CU mode decision algorithm are described in section II. Experimental results are provided in section III, and section IV concludes this paper.

II. THE PROPOSED PERCEPTUAL CODING ALGORITHM

Both of the proposed perceptual coding scheme and fast mode decision algorithm are based on VSM. To obtain VSM, a classic VS detection model proposed by Seo [9] is used.

A. VS Detection Model

Seo [9] proposed a novel unified framework for both static and space-time saliency detection. The proposed method is practically appealing because it is nonparametric, fast, and robust to uncertainty in the data. The overall algorithm is divided into two steps: local data structure capture and self-resemblance measure.

A local steering kernel (LSK) based method is used to capture local data structure. The LSK is defined as,

This work was supported by Fundamental Research Funds for the Central Universities (3102014JCQ01057).

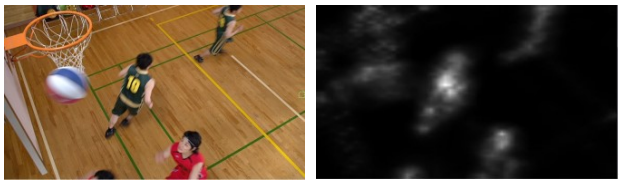
$$K(x_l - x_i) = \frac{\sqrt{\det(C_l)}}{h^2} \exp\left\{\frac{(x_l - x_i)^T C_l (x_l - x_i)}{-2h^2}\right\} \quad (1)$$

where h is a global smoothing parameter, and C_l is a covariance matrix which can be estimated from gradient vectors within local analysis window. The size and shape of LSK are changed adaptively by C_l according to the texture. So LSK can capture geometric construction in an image or video exceedingly well.

The self-resemblance measure is derived from a locally data-adaptive kernel density estimator which makes the saliency detection algorithm more effective and simpler than other methods and does not require any training. The likelihood of saliency can be calculated as

$$S_i = \frac{1}{\sum_{j=1}^{N} \exp\left(\frac{-1 + r(F_i, F_j)}{s^2}\right)} \quad (2)$$

where F is feature matrix composed of LSKs, $\rho(F_i, F_j)$ is matrix cosine similarity between two feature matrices. The saliency in every pixel makes up a VSM which indicates the distribution of saliency in a frame.



(a) Original frame (b) Visual saliency map

Fig. 1. A frame from BasketballDrill and its visual saliency map

An original frame from test sequence BasketballDrill and its corresponding VSM by the above model are shown in Fig .1. In a VSM, the lighter a pixel is the more attention the audience will pay. In perceptual video coding VSM is used as the direction to control distortion in perceptual coding. What's more, VSM can reflect the complexity of texture and movements in frames in some degree as geometric construction is used as features in the above VS detection model.

B. Refined Distortion Allocating Scheme

In general, there are mainly three types of distortion control methods, in pixel level, transformed coefficient level and CU level, respectively. Our scheme is implemented in CU level, which is easy to be realized and will not cause any extra computation burden to the encoder compared with the other two.

As CUs with high saliency will attract more attention, CUs with low saliency can tolerate more distortion. In this paper, distortion in a CU is allocated according to the average saliency. More distortion is introduced to CUs with low saliency while less to those with high saliency. The amount of distortion in a CU is determined by JND related method. For simplicity, CUs are classified into five categories according to the average saliency. Each category corresponds to a distortion level. The average saliency of a CU is

$$Avg_N = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} S_i(i,j) / N^2 \quad (3)$$

where N is CU size, and $S_i(i, j)$ is the saliency of the pixel with coordinate (i, j) in a CU. The average saliency of a CU in the following paper can also be calculated by (3).

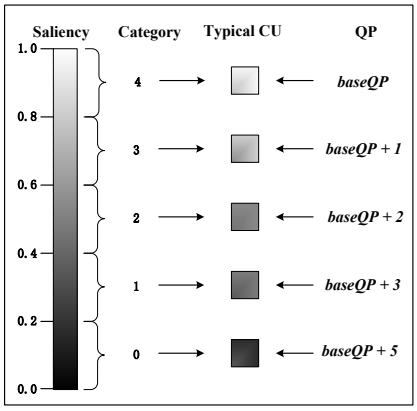


Fig. 2. Mapping from saliency to QP

In video coding, QP is used to adjust distortion. The larger a QP is, the more distortion will be introduced. For each CU category, an extra value is added to the original QP to control distortion. The value of deltaQP is determined by JND. As shown in Fig.2, the value of 1 is assigned to the deltaQP of category 4, and subjective quality experiment is conducted to test whether any noticeable distortion will be caused. If no, the deltaQP will be added until some distortion can be noticed. Then deltaQP-1 will be added to the original QP to get an extra bit-rate reduction. The deltaQP for CUs in other category can be obtained in the same way. The mapping relationship between the average saliency of CUs and the corresponding QP is shown in Fig. 2, where baseQP is the original QP of a CU. It is clear that the QP assigned to a CU is in inverse proportion to its average saliency. That's to say more distortion is introduced to CUs with low saliency while less to CUs with high saliency.

C. Fast CU Mode Decision Algorithm

A pyramid CU mode decision scheme is used in HEVC standard, as shown in Fig. 3. In the pyramid CU mode decision scheme, a CTU is recursively split from depth 0 to depth 3. There exists a lot of computing redundancy in such an enumeration-based mode decision structure.

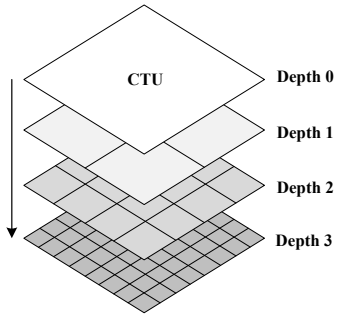


Fig. 3. Pyramid-like mode decision structure

Usually, the maximal depth of optimal CU mode is closely related to texture and movements complexity in frames. A CU of high complexity tends to be split into smaller CUs to achieve more precise motion estimation. On the contrary, mode decision in small depth for CUs of low complexity is not as important as the one in big depth. That is to say mode decision in some depth can be skipped according to the complexity of a CU.

As described in section II.A, spatial and time texture are used as features in Seo's [9] visual saliency model. In other words, this visual saliency model can reflect texture and movements complexity in some degree. So, visual saliency by Seo [9] can be used as an estimation of complexity. Complexity of CUs is divided into four grades according to the average saliency in this paper. The saliency range of each grade and the corresponding CU depth needed to be checked are determined by counting the average VS of CTUs. For example, when the maximum depth is 0, the average VS of CTUs is calculated, and then saliency range of depth 0 can be obtained. The saliency range of other depth can be obtained in a similar way. There are crossing areas among these saliency ranges. Finally, the midpoints of the crossing areas are selected as dividing points of each grade. The mapping relationship between the average VS and CU depth needed to be checked by the above method is shown in Fig. 4.

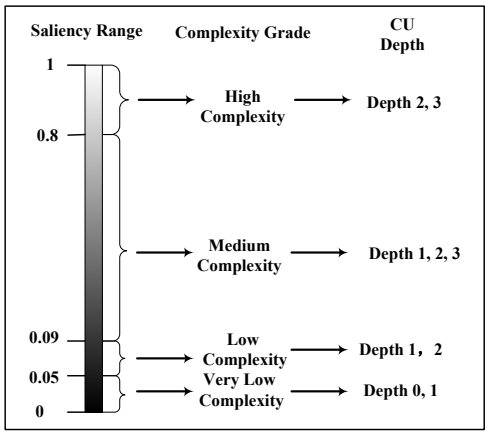


Fig. 4. Mapping from saliency to CU depth

CU depth can be predefined by the fast mode decision method described above. But there is a situation where pixels with high saliency concentrate in the corner of a CU, and the quality of these pixels can not be strengthened as pixels in a CU share the common QP. So it is necessary to split such a CU into small CUs. For a CTU or a 32×32 CU, the average VS of their son CUs is used to determine whether to split directly. The process can be described as,

$$if \quad max(Avg^k_N) - min(Avg^k_N) > \Upsilon, \quad then\ skip\ depth\ i \qquad (4)$$

where Avg^k_N is the average visual saliency of one of the four son CUs in a $N\times N$ CU, k is the index of the son CUs ranging from 0 to 3, i is the depth of $N\times N$ CU, and Υ is the corresponding threshold (assigned 0.2 in our experiment). Avg^k_N can be calculated by (3) with N being 32 or 16.

In general, the proposed perceptual coding algorithms are shown in Fig. 5. Information used to control HEVC encoder is produced by the proposed algorithms. Refined distortion allocating method and fast mode decision are totally based on these control information. It should be noted that the proposed method is fully compliant with the HEVC standard.

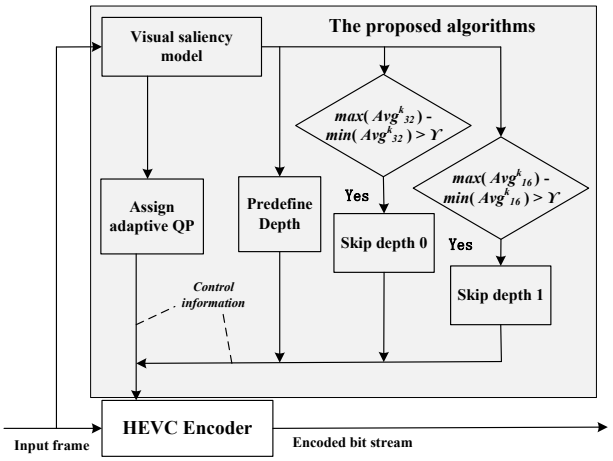


Fig. 5. Proposed perceptual coding algorithms

III. EXPERIMENTAL RESULT

The proposed algorithms are implemented into HEVC test model HM13.0 to verify its efficiency. Then subjective quality tests are conducted to evaluate the performance of the proposed perceptual coding scheme. The protocol we used for the tests is the Double-Stimulus Continuous Quality Scale (DSCQS) protocol recommended in Rec. ITU-R BT.500 [10]. Before testing, the assessors were carefully introduced about the assessment procedure and method. HEVC test sequences were used in the tests. Each sequence was coded with four quantization parameter values: 22, 27, 32, and 37, respectively.

To assess the subjective quality differences among Kim [3], Li [5] and the proposed algorithm for each test sequence, the Difference Mean Opinion Score (DMOS) values are calculated as $DMOS = MOS_{perceptual\ coding} - MOS_{HM13.0}$. The smaller the absolute DMOS values are, the closer the subjective qualities are to those of the original HM13.0. The subjective quality test results are shown in Table 1. It can be seen from Table 1 that visual quality by the proposed algorithm is not inferior to HM13.0 as the absolute DMOS value is even smaller than 0.1. Visual quality by the proposed algorithm is also similar to Li [5], as both of the absolute DMOS values are very small and similar. But up to 10% more bit-rate can be saved by the proposed algorithm compared with Li [5]. The algorithm in Kim [3] is more efficient for high bit-rate video and more bit-rate can be reduced by the proposed algorithm for low bit-rate.

The efficiency of our fast mode decision algorithm is shown in Table 2, where VS time is the percentage of the time used to calculate VSM. As we can see from Table 2, the proposed algorithm is even more efficient than some traditional fast algorithm [7] and 43.48% encoding time can be saved. The time used to calculate VSM is only 1.87% on average which is insignificant compared with the whole encoding time.

IV. CONCLUSION

A perceptual video coding scheme with refined distortion allocating method and a fast mode decision algorithm are proposed in this paper. Refined distortion allocating method is based on visual saliency in CU level by adaptive QP. In the proposed fast mode decision algorithm, saliency is also used as an approximation of texture complexity and movements. As shown in subjective tests, the proposed perceptual coding scheme achieves up to 22.52% bit-rate reduction with negligible perceptual quality loss. In addition, 43.48% computation complexity in encoder can be reduced by the proposed algorithm.

REFERENCES

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

[2] Chen, Zhenzhong, and C. Guillemot. "Perceptually-Friendly H.264/AVC Video Coding Based on Foveated Just-Noticeable-Distortion Model," *Circuits & Systems for Video Technology*, *IEEE Transactions on*, vol. 20, no. 6, pp. 806-819, June 2010.

[3] Kim, J., S. Bae, and M. Kim. "An HEVC-Compliant Perceptual Video Coding Scheme based on JND Models for Variable Block-sized Transform Kernels," *Circuits & Systems for Video Technology* , *IEEE Transactions on*, vol. 25, no. 11, pp. 1786-1800, November 2015.

[4] Gupta, Rupesh, M. T. Khanna, and S. Chaudhury. "Visual saliency guided video compression algorithm," *Signal Processing Image Communication*, vol. 28, no. 9, pp. 1006-1022, 2013.

[5] Li, Y., Liao, W., Huang, J., He, D., and Chen, Z. "Saliency based perceptual HEVC," *Multimedia and Expo Workshops (ICME)*, *2014 IEEE International Conference on*, 2014.

[6] Wang, H., Wang, L., Hu, X., Tu, Q., and Men, A. "Perceptual video coding based on saliency and Just Noticeable Distortion for H.265/HEVC," *Wireless Personal Multimedia Communications (WPMC)*, *2014 International Symposium on*, pp. 106-111, 2014.

[7] Shen, L., Liu, Z., Zhang, X., Zhao, W., & Zhang, Z. "An effective cu size decision method for hevc encoders," *Multimedia*, *IEEE Transactions on,* vol. 15, no. 2, pp. 465-470, 2013.

[8] Vanne, Jarno, Marko Viitanen, and Timo D. Hamalainen. "Efficient mode decision schemes for HEVC inter prediction," *Circuits and Systems for Video Technology*, *IEEE Transactions on*, vol. 24, no. 9, pp. 1579-1593, 2014.

[9] Seo, Hae Jong, and Peyman Milanfar. "Static and space-time visual saliency detection by self-resemblance," *Journal of vision*, vol. 9, no. 12, pp. 74-76, 2009.

[10] ITU-T. BT500-11, "Methodology for the subjective assessment of the quality of television pictures," 2002

TABLE I. RESULT OF DSCQS TEST

sequence	QP = 22					QP = 27				
	Kim[3]	Li[5]		proposed		Kim[3]	Li[5]		proposed	
	bit-rate saving(%)	*bit-rate saving(%)*	*DMOS*	*bit-rate saving(%)*	*DMOS*	*bit-rate saving(%)*	*bit-rate saving(%)*	*DMOS*	*bit-rate saving(%)*	*DMOS*
Cactus	49.1	—	—	33.38	-0.063	19.67	—	—	19.26	0.125
Kimono	6.81	12.85	0	20.45	0	5.12	10.04	-0.133	19.07	0
ParkScene	34.11	13.64	0.188	22.71	0.063	23.86	10.62	-0.2	20.24	-0.125
FourPeople	—	—	—	24.62	0	—	—	—	11.65	-0.063
KristenAndSara	—	—	—	23.31	-0.063	—	—	—	17.98	-0.063
BQMall	26.49	13.64	0.188	17.03	-0.438	18.35	7.81	0.267	13.01	-0.063
PartyScene	31.00	11.65	0.125	17.54	0.250	25.03	9.25	0.067	14.83	-0.188
RaceHorses	37.02	10.24	-0.438	21.08	0	25.91	8.12	-0.133	15.63	0.063
Average	30.76	12.40	0.013	22.52	-0.031	19.66	9.17	-0.026	16.46	-0.039
sequence	**QP = 32**					**QP = 37**				
	Kim[3]	**Li[5]**		**proposed**		**Kim[3]**	**Li[5]**		**proposed**	
	bit-rate saving(%)	***bit-rate saving(%)***	***DMOS***	***bit-rate saving(%)***	***DMOS***	***bit-rate saving(%)***	***bit-rate saving(%)***	***DMOS***	***bit-rate saving(%)***	***DMOS***
Cactus	5.47	—	—	14.77	-0.063	3.9	—	—	12.40	0
Kimono	0.91	9.26	0.133	13.42	-0.125	3.07	8.76	0.133	13.44	-0.125
ParkScene	9.49	8.85	-0.267	17.85	0.063	7.02	8.26	-0.533	15.38	-0.188
FourPeople	—	—	—	5.32	0	—	—	—	1.34	0
KristenAndSara	—	—	—	7.08	-0.063	—	—	—	1.66	0
BQMall	6.25	6.38	0.133	9.36	-0.312	5.01	5.9	0.2	5.88	-0.312
PartyScene	11.91	8.22	0.133	11.62	0.312	6.66	8.15	-0.267	10.82	0.25
RaceHorses	7.22	7.26	-0.2	13.04	-0.063	7.31	6.81	-0.067	11.93	-0.188
Average	6.88	8.0	-0.014	11.56	-0.031	5.50	7.58	-0.107	9.01	-0.07

TABLE II. RESULT OF COMPLEXITY REDUCTION

	sequence	Cactus	Kimono	Park Scene	Four People	Kristen AndSara	BQMall	Party Scene	Race Horses	Average
Shen[7]	**time saving(%)**	43	41	41	—	—	42	32	20	36.5
proposed	**time saving(%)**	46.42	46.07	46.21	49.1	50.7	36.52	36.48	36.37	43.48
	VS time(%)	0.83	0.63	0.8	2.44	2.44	3.03	2.67	2.14	1.87