2013 Sixth International Conference on Advanced Computational Intelligence
October 19-21, 2013, Hangzhou, China

A ROI-based Bit Allocation Scheme for HEVC towards Perceptual Conversational Video Coding

Xin Deng, Mai Xu, and Zulin Wang

***Abstract*—In this paper, we put forward a simple yet efficient scheme, based on the HEVC platform, to improve the subjective quality of conversational videos. Our scheme, in fact, is a perceptual bit allocation scheme (PBAS) for HEVC, aiming at increasing the amount of bits assigned in ROI regions. First, PBAS scheme employs the robust SURF cascade face detector to extract the regions of interest (ROI) in a conversational video. Then, based on the extracted ROI regions, a perceptual rate-distortion model is presented in PBAS scheme towards subjective quality improvement. By optimizing the rate-distortion model, we can have the optimal settings of bit rates for the ROI and non-ROI regions using PBAS scheme, respectively. Finally, the experimental results show that the PBAS scheme can improve significantly the subjective quality of conversational videos compressed by HEVC.**

I. Introduction

NOWADAYS,with the ever-increasing demands for smart phones, the traditional mobile phones have no choice but to step down from the stage of history. Instead of them, iPhone and Android, begin to emerge fleetly, showing their dominant positions. Due to the great success of smart phones, plenty of conversational video products, such as FaceTime, are flooding into our lives, facilitating the visual communications of human. However, the huge amount of video data demands adequate storage capacities to store and sufficient bandwidth to transmit. All these changes and demands are calling for a new super video coding design to come out. Fortunately, High Efficiency Video Coding (HEVC) standard [1], also called H.265, has been announced to be formally established. With flexible picture partitioning, parallel encoding/decoding and some other cutting-edge technologies, HEVC has eminent compression performance, much more better than the preceding H.264/AVC [2].

One of most important tasks of conversational video coding, such as HEVC, is the bit allocation, which aims at finding some improved balance between video quality and bit rate. Unfortunately, as evaluated in this paper, the bit allocation scheme in HEVC is not adaptive to subjective video quality, which is of vital importance to improve the quality of human experience (QoE). Therefore, we propose in this paper a simple yet effective bit allocation scheme (namely PBAS), upon region of interest (ROI) detected by [10], to improve the subjective quality of conversational video coding of HEVC. In this paper, subjective quality refers to the video quality perceived and judged by humans, the same as [4].

Towards the improvement of subjective video quality, the bit allocation and rate control, in earlier video coding standards, have been extensively studied in [4] [5] [6] [7] [8] over one decade. More specifically, for H.263, a perceptual rate control scheme [4] was proposed to vary the quality parameters (QP), which is dependent on perceptual sensitive weight map. Then, this scheme can magnify the importance of the skin and sensitive regions in a conversational video. For H.264, a ROI-based bit allocation method [6] was proposed to optimize the QPs in conversational video coding, constrained by the exploited skin-tone information. For the survey of perceptual video coding, see [7].

However, to our best knowledge, the perceptual performance of HEVC, in particular for conversational video coding, has yet to be analyzed. Meanwhile, the bit allocation for the perceptual video coding remains to be done in HEVC. Therefore, the main contributions of this paper is listed as follows:

(1) *Evaluation of perceptual performance of HEVC.* The same as [4] we consider the ROI regions in the conversational video being faces, extracted by state-of-the-art face detector [10]. Then, on the basis of the detected faces in the video, we evaluate the perceptual performance, by plotting the PSNRs of detected ROI, non-ROI and whole regions in the conversational video, at various bit-rates.

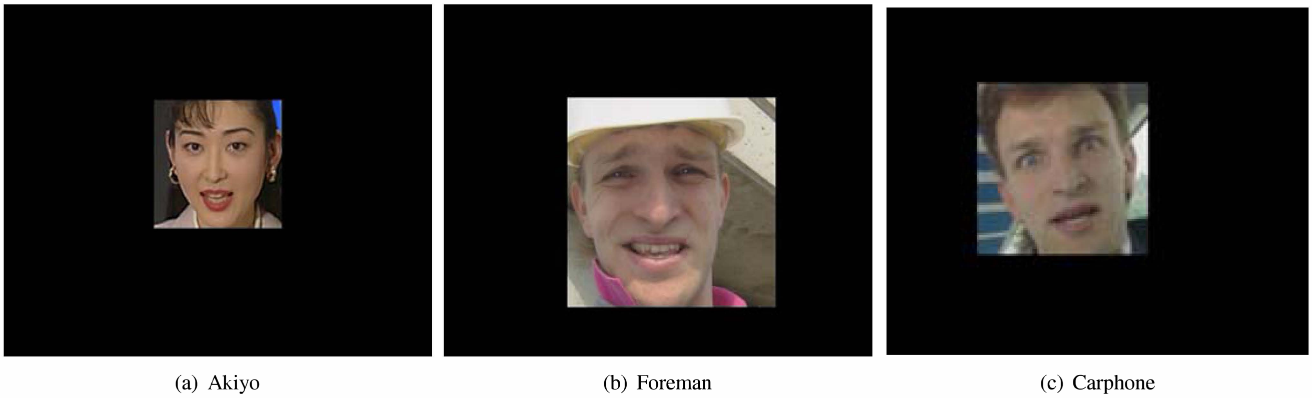
(2) *Bit allocation scheme.* We aim at improving the subjective perceptual performance of the conversational video. Towards this end, PBAS scheme is proposed to allocate more bits to ROI regions of the video, by optimizing a novel subjective rate-distortion model. Instead of setting various QPs to each macroblock as the existing perceptual video coding approaches[4] [6], only bit-rates need to be configured for different regions based on ROI, thus enjoying the simplicity of small modification of the conventional rate control scheme in HEVC.

The rest of this paper is organized as follows: Section II presents an analysis on the perceptual performance of conversational video coding of HEVC, which shows the necessariness of our work. Then, a bit-allocation scheme is proposed for conversational video coding of HEVC, with a subjective rate-distortion model in Section III. In addition, some experimental results are presented in Section IV to validate the proposed PBAS scheme. Finally, Section V concludes this paper.

Xin Deng, Mai Xu, and Zulin Wang are with the Department of Electronic and Information Engineering, Beihang University, Beijing, 100191, China.

This work was supported by the NSFC project 61202139

978-1-4673-6343-3/13/$31.00 ©2013 IEEE



(a) Akiyo

(b) Foreman

(c) Carphone

Fig. 1. The detected faces in one frame of conversational video, using SURF cascade face detector [10]. Here, the rectangular region is used to include each detected face.

II. Evaluation on perceptual performance of conversational video coding of HEVC

A. *Face detection for extracting perceptual ROI regions*

The perceptual mechanism of human visual system (HVS) has a long way to go yet. Thus, the precise identification of ROI regions, indeed, is intractable in the perceptual video coding. Fortunately, we have an important cue for perceptual model in conversational video communications, i.e. extracting faces as the ROI regions. Faces were widely used as the ROI regions in a lot of perceptual video coding approaches [4] [6] [9]. However, all these approaches utilized the skin tone as face detector, which is not robust to various lighting conditions and other objects with similar colors. In this paper, we apply a robust speed up robust features (SURF) cascade face detector [10], benefiting from the most recent success on computer vision, to extract the ROI regions for our conversational video coding.

First, the SURF [11] has to be extracted as the input to the face detector. Then, given SURF features $\mathbf{x}_k$ of kth local patches, a logistic regression model can be applied to define a weak classifier:

$$c(\mathbf{x}_k) = P(y_k = \pm 1 | \mathbf{w}, \mathbf{x}_k) = \frac{1}{1 + \exp(-y_k \mathbf{w}^T \mathbf{x}_k)}, \quad (1)$$

where $\mathbf{w}$ is the parameter of model, leant from training data; $y_k = \pm 1$ is the label of patch, indicating whether the patch belongs to the face or not. Next, considering $c(\mathbf{x}_k)$ as a base learner, we can obtain a strong classifier (i.e. a linear combination of $c(\mathbf{x}_k)$) using a boosting procedure with the boosting rounds being $t = \{1, 2, \ldots, T\}$:

$$C^t(\mathbf{x}) = \max_{k=\{1,2,\cdots,K\}} J(C^{t-1}(\mathbf{x}_k) + \alpha_k c(\mathbf{x}_k)), \quad (2)$$

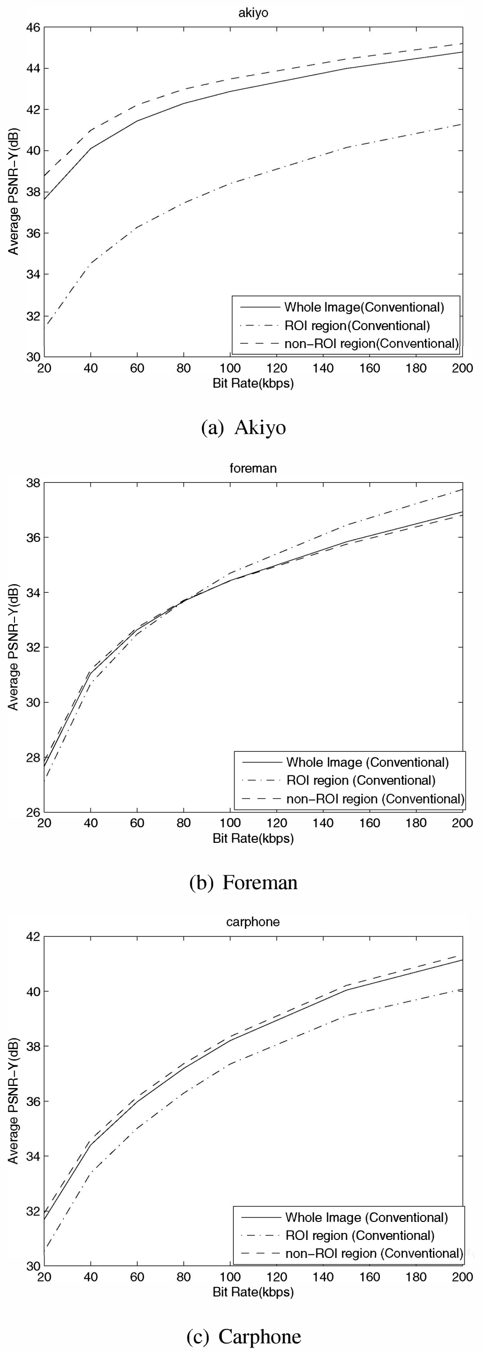
where α_k is the weight of base learner $c(\mathbf{x}_k)$ ranging from 0 to 1, and $J(\cdot)$ is AUC score (Area under ROC curve). Finally, the face detector may output the patches $C^t(\mathbf{x}) > \theta_i$ as the face regions, in which θ_i is a threshold. For more details, see [10].

Figure 1 displays the detected faces in three standard testing sequences of conversational videos. As seen from this figure, the faces in three testing sequences are able to be extracted effectively, thus yielding the ROI regions for perceptual video coding. On the basis of the detected faces in each video frame, we show in the next subsection the rate-distortion performance of HEVC over ROI, non-ROI and whole regions for various standard testing sequences.

B. *Analysis of perceptual performance on HEVC for conversational scenario*

Admittedly, the latest video coding standard, HEVC standard, has improved a lot in comparison with the preceding H.264/AVC. However, it still has some undesirable defects. As we all know, the HEVC adopts a similar coding structure as H.264/AVC. This results in a challenge resembling the H.264/AVC, that is the ill-suited bit allocation problem which may cause an unsatisfactory visual perception performance. In order to explain this issue better, we now test the rate-distortion characteristic of three standard conversational video sequences, compressed by HEVC. The test sequences we choose are Akiyo, Foreman and Carphone. In each video sequence, we examine the changes of PSNR under a range of bit rates to illustrate the rate-distortion performance. The rate-distortion curves of the test sequences are plotted in Figure 2. The PSNR$_W$, PSNR$_F$, and PSNR$_B$ are assigned to denote the PSNRs of the whole image, the ROI regions, and the non-ROI regions, which have been split using the method presented in Section II-A.

It can be seen from Figure 2 that the average PSNR$_F$ is lower than the PSNR$_B$ for nearly all the three sequences except the Foreman sequence, where the PSNR$_F$ is almost the same as the PSNR$_B$. However, generally speaking, it is the ROI regions that we pay more attention to. While the non-ROI regions usually cannot catch our eyes so much. Hence, the reasonable approach is higher the PSNRs in ROI regions and lower the PSNRs in non-ROI regions. Unfortunately, from Figure 2, we can conclude that this target is not well achieved by HEVC. Since it does not take into consideration the unequal perceptual weights between the ROI and non-ROI regions, when allocating the bits in conversational video coding. Obviously, this issue may result in undesirable



(a) Akiyo

(b) Foreman

(c) Carphone

Fig. 2. The rate-distortion performance of three conversational video sequences using conventional rate control scheme compressed by HEVC.

performance, as it runs counter to the HVS. Therefore, it is necessary to look for an efficient way, which can allocate the bits more reasonably, in order to be consistent with the HVS.

III. Bit allocation scheme for conversational video coding of HEVC

In this section, we first present in Section III-A a perceptual rate-distortion model, according to HVS. Then, in Section III-B PBAS scheme is proposed to optimize the perceptual rate-distortion model, thus being able to improve the sujective quality of conversational video coding.

A. Perceptual rate-distortion model for conversational video coding

The tiles, as a new characteristic of HEVC, enable the parallel process of video coding. Therefore, the ROI and non-ROI regions can be compressed using different tiles of HEVC separately, subject to different distortions. Here, we utilize the simple rate-distortion model [12], but with different optimization objectives of distortion over ROI and non-ROI regions, for improving the subjective perceptual performance in ROI regions.

To be more specific, first of all, the conversational video scenes are required to be split into two parts, the ROI and the non-ROI regions, by employing the face detector introduced in Section II-A. The amounts of bits allocated to ROI and non-ROI regions are denoted by B_I and B_o, respectively. With the rate-distortion model put forward in [12], we can model B_I and B_o as follows:

$$B_I = A_I(\frac{K\sigma_I^2}{Q_I^2} + C) \quad \text{and} \quad B_o = A_o(\frac{K\sigma_o^2}{Q_o^2} + C), \qquad (3)$$

where σ_I and σ_o are the standard deviations of pixel values in ROI and non-ROI regions; A_I and A_o represent the number of pixels in ROI and non-ROI regions; Q_I and Q_o denote the quantization parameters of HEVC in ROI and non-ROI regions. In addition, K and C in Equation (3) are constants. As the value of C is much smaller than that of $\frac{K\sigma_I^2}{Q_I^2}$ or $\frac{K\sigma_o^2}{Q_o^2}$, we simply set C to be 0.

In order to take full account of the subjective perceptual performance according to HVS, we introduce D_I and D_o as the perceptual distortions of ROI and non-ROI regions in the following:

$$D_I = w^2(\frac{Q_I^2}{12}) \quad and \quad D_o = \frac{Q_o^2}{12}, \qquad (4)$$

where w indicates the relative importance of ROI regions over non-ROI regions. The larger w is set, the greater significance ROI region is endowed. w=1 implies that the two regions are of the same importance, for non-perceptual case.

Finally, Equations (3) and (4) determine the perceptual rate-distortion model used for the proposed PBAS scheme in the next subsection.

B. The proposed perceptual bit allocation scheme

Based on the aforementioned perceptual rate-distortion model, the problem of perceptual bit allocation can be reduced to the following optimization problem:

$$\min_{B_I}(\frac{KA_I w^2\sigma_I^2}{12B_I} + \frac{KA_o\sigma_o^2}{12B_o}) \quad \text{s.t.} \quad B_I + B_o = B. \qquad (5)$$

In Equation (5), B is the target bit-rate we have for the conversational video compressed by HEVC. The optimization problem of Equation (5) can be solved by a simple deviation.

Then, given target target bit-rate B, optimal B_I of the PBAS scheme can be obtained by

$$B_I = \frac{1}{1+\frac{\sigma_o}{w\sigma_I}\sqrt{\frac{N}{A_I}-1}}B, \tag{6}$$

where N=A_I+A_o, denotes the total number of pixels in each frame. Furthermore, once obtaining B_I, B_o can be computed easily by $B_o = B - B_I$.

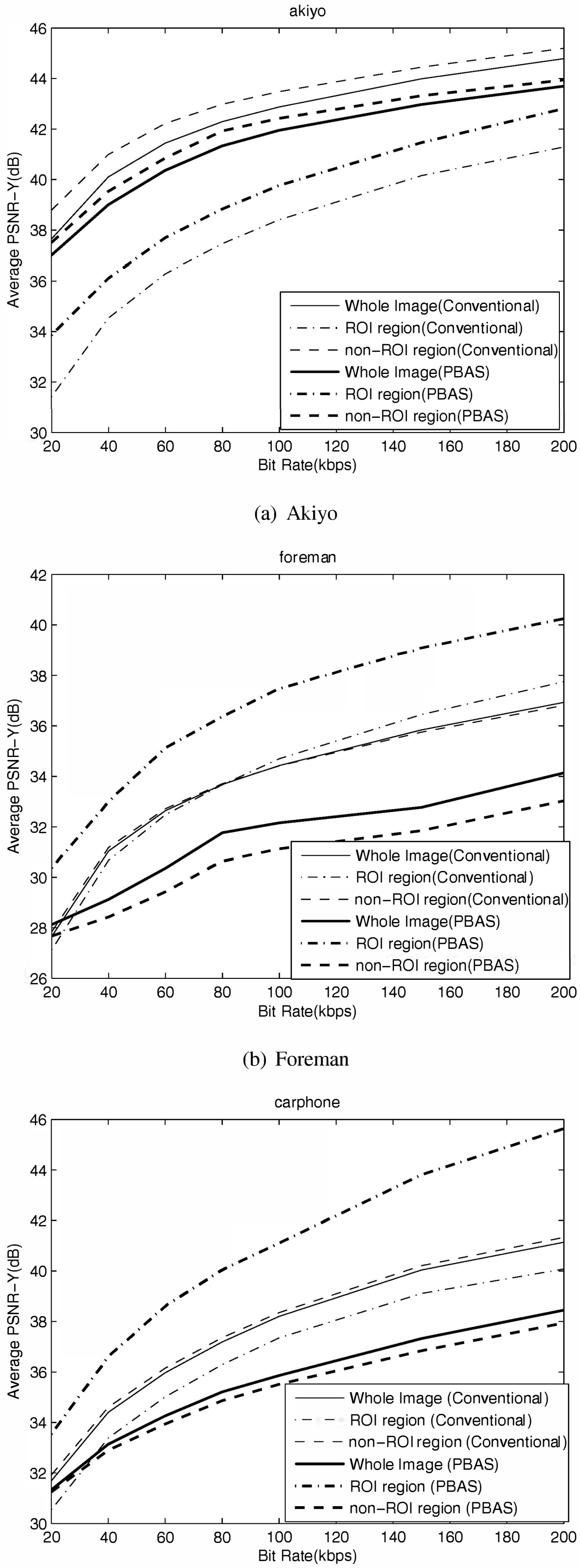
From Equation (6), we can see that B_I is depended on B and w, when given a conversational video, where σ_o, σ_I, A_I and A_o are its internal factors. In other words, once B and w are fixed, B_I can be computed using the result of Equation (6). Note that the larger w is, the greater B_I is, indicating the distortion improvement over ROI regions. However, such improvement is at the expense of the worse distortion in non-ROI regions. Let us imagine an extreme case where w tends to infinity, then B_I can be seen as the total target bit-rate B. This analysis result also makes sense, since there is no attention paid to non-ROI regions, the ROI regions are deserved to be allocated all the bit rates. Hence, we can see that our PBAS scheme is advisable and feasible. What we need to do is adjusting the value of w that implies the relative importance of ROI regions, to obtain the subjective video quality we want. The specific adjustment of w and its influence on the rate-distortion performance are discussed using some experiments in the next section.

IV. Experimental results

The proposed PBAS scheme has been implemented in three test sequences: Akiyo, Foreman and Carphone. In our experiments, we compare the rate-distortion performance between the proposed scheme and the conventional rate control scheme with HEVC at various bit rates. The results of our experiments are shown in Figure 3.

From Figure 3, we can easily see that the PSNRs of ROI regions in all the three sequences are greatly enhanced by the PBAS scheme. Specifically, for the Foreman and Carphone sequences, under the PBAS scheme, the PSNRs of ROI regions are higher than the PSNRs of non-ROI regions. For the Akiyo sequence, although the PSNR of ROI region is still lower than that of non-ROI region, the gap between the two regions is tremendously narrowed. We also show in Figure 4 the 47th reconstructed frame of Foreman sequence, which is compressed by HEVC, with conventional rate control and the PBAS schemes respectively, at three different bit rates. From this figure, we can notice that the quality of ROI regions is greatly improved by PBAS scheme. Furthermore, from the comparison of three pairs ((a)-(b) and (c)-(d) and (e)-(f)) shown in Figure 4 with different bit rates, we can also conclude that our PBAS scheme works better at lower bit rate.

Our aim is to improve the quality of ROI regions, and as the cost, the PSNRs of non-ROI and the whole regions are reduced. However, as we have mentioned before, thanks to the HVS, most of our attention are concentrated on the ROI regions, while little eyesight stays in the non-ROI regions

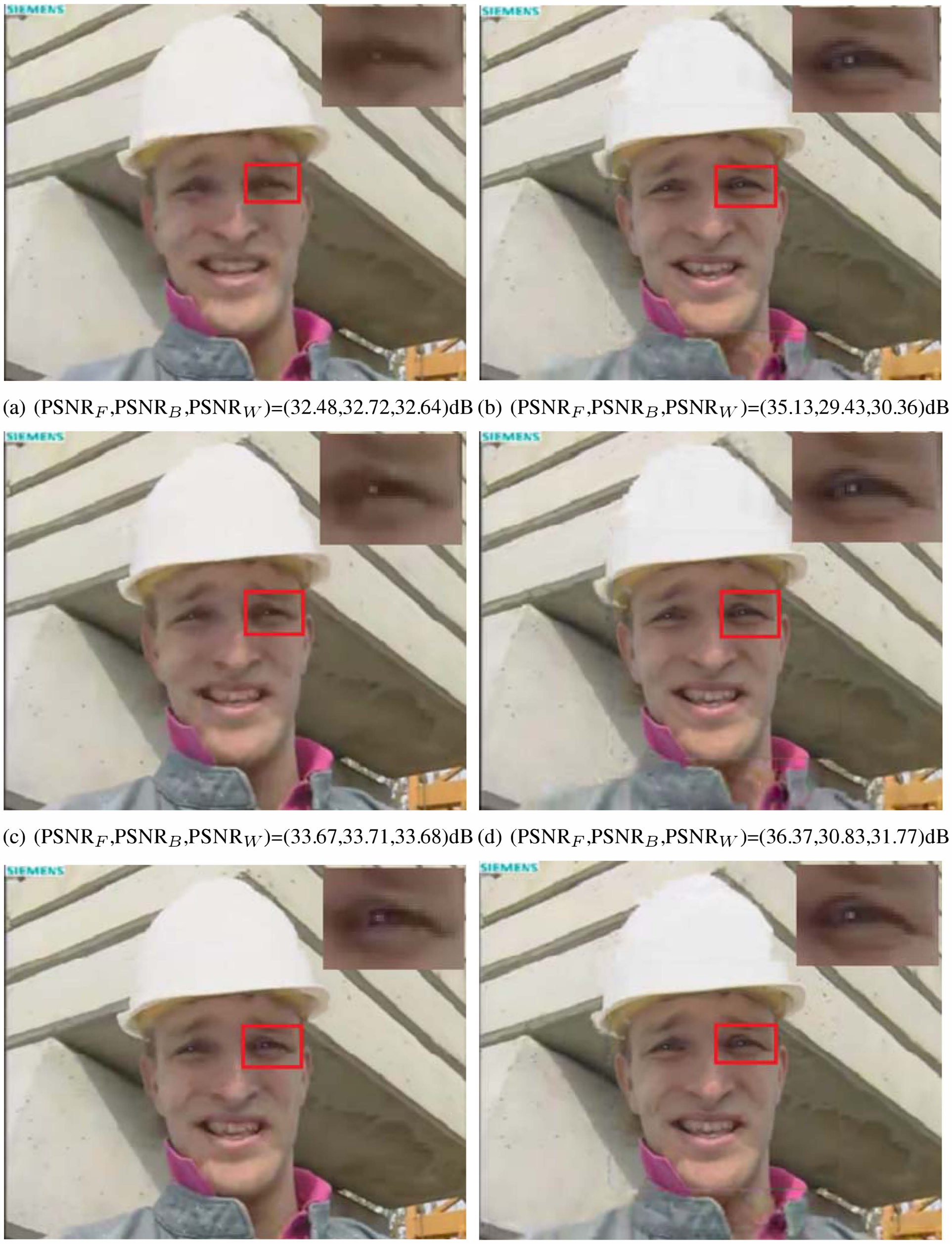


(a) Akiyo

(b) Foreman

(c) Carphone

Fig. 3. The rate-distortion performance comparison between conventional and the proposed PBAS schemes in HEVC for ROI, non-ROI and whole regions in three sequences



(a) (PSNR$_F$,PSNR$_B$,PSNR$_W$)=(32.48,32.72,32.64)dB (b) (PSNR$_F$,PSNR$_B$,PSNR$_W$)=(35.13,29.43,30.36)dB

(c) (PSNR$_F$,PSNR$_B$,PSNR$_W$)=(33.67,33.71,33.68)dB (d) (PSNR$_F$,PSNR$_B$,PSNR$_W$)=(36.37,30.83,31.77)dB

(e) (PSNR$_F$,PSNR$_B$,PSNR$_W$)=(34.70,34.42,34.43)dB (f) (PSNR$_F$,PSNR$_B$,PSNR$_W$)=(37.48,31.13,32.16)dB

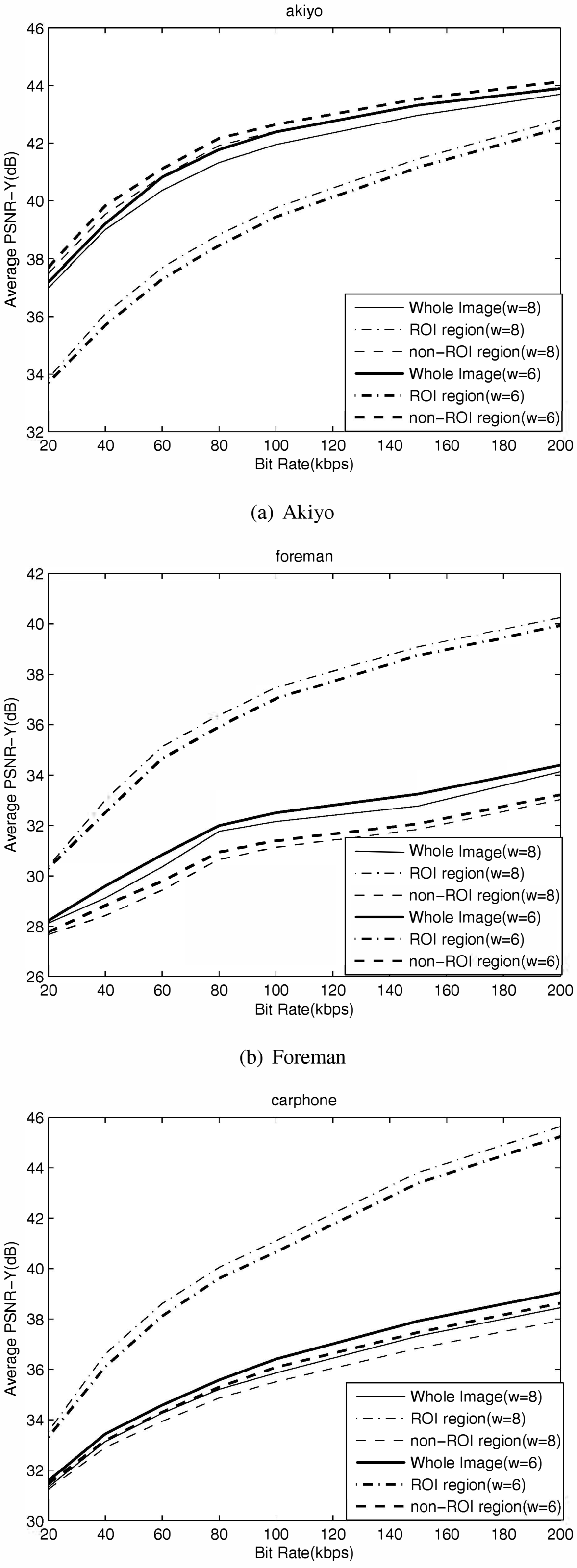
Fig. 4. The 47th decoded frame for Foreman at various bit rates: (a) Conventional scheme at 60kbps, (b) PBAS at 60kbps, (c) Conventional scheme at 80kbps, (d) PBAS at 80kbps,(e) Conventional scheme at 100kbps, (f) PBAS at 100kbps.

long. Hence, the reduction of PSNRs in non-ROI regions cannot set off a huge tempest for the integral subjective video quality perception. As we can see from Figure 4, the decrease of PSNRs in non-ROI regions has little impact on the whole video quality.

Figure 5 shows the comparison of rate-distortion performance using PBAS scheme with different w (set to 8 and 6 in this experiment). Obviously, when w is adjusted from 8 to 6, the PSNRs of ROI regions become lower, while the PSNRs of non-ROI and whole regions turn higher. This keeps in accordance with the theoretical analysis discussed in section III. That is, if we want to put more emphasis on the ROI regions, we only need to increase the value of w, and vice versa.

V. Conclusions

In this paper, we have proposed a perceptual bit allocation scheme (PBAS), for improving the subjective visual quality of conversational video on the HEVC platform. Although the latest video coding standard, HEVC, has employed plenty of novel techniques to improve the video compression performance, it does not take full account of the unequal weights between ROI and non-ROI regions, thus resulting in



(a) Akiyo

(b) Foreman

(c) Carphone

Fig. 5. The comparison of rate-distortion performance using PBAS scheme when w is set to 8 and 6, for ROI, non-ROI and whole regions in three sequences

a relatively low quality for ROI regions. In order to make it consistent with HVS, we need to take some measures to improve the video quality of ROI regions. The most effective method is allocating more bits to ROI regions, seen as the motivation of our scheme.

In our PBAS scheme, the ROI regions are extracted from the conversational video, using the SURF cascade face detector. Then, based on the extracted ROI regions, a perceptual rate-distortion model is presented. By optimizing the rate-distortion model, we can obtain the different optimal bit rates set for ROI and non-ROI regions, respectively. Our experimental results on compressing the standard conversational videos with HEVC have verified that the PBAS scheme performs better in terms of the PSNR in ROI regions, against the conventional rate control scheme of HEVC.

References

[1] G. J. Sullivan, J. Ohm, and W. Han, "Overview of the High Efficiency Video Coding (HEVC) standard", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, Dec. 2012.

[2] T. Wiegand, G. J. Sullivan, and G. Bjontegaard, "Overview of the H.264/AVC video coding standard", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, Jul. 2003.

[3] J. Li, T. Wang, and Y. Zhang, "Face detection using SURF cascade", *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011.

[4] X. Yang, W. Lin, Z. Lu, X. Lin, S. Rahardja, E. Ong, and S. Yao, "Rate control for videophone using local perceptual cues", *IEEE Transations on Circuits and Systems for Video Technology*, vol. 15, no. 4, Apr. 2005.

[5] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compression with optimal rate control", *IEEE Transations on Image Processing*, vol. 10, no. 7, Jul. 2001.

[6] Y. Liu, Z. G. Li, and Y. C. Soh, "Region-of-interest based resource allocation for conversational video communication of H.264/AVC", *IEEE Transations on Circuits and Systems for Video Technology*, vol. 18, no. 1, Jan. 2008.

[7] J. Lee and T. Ebrahimi, "Perceptual video compression: a survey", *IEEE Journal of Selected Topics in Signal Processing*, vol. 18, no. 1, Jan. 2012.

[8] Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable distortion model", *IEEE Transations on Circuits and Systems for Video Technology*, vol. 20, no. 6, Jun. 2010.

[9] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, "Low bit-rate coding of image sequences using adaptive regions of interest", *IEEE Transations on Circuits and Systems for Video Technology*, vol. 8, no. 8, Dec. 1998.

[10] H. Baya, A. Essa, T. Tuytelaarsb, and L. V. Goola, "Speeded-Up Robust Features (SURF)", *Computer Vision and Image Understanding*, vol. 110, no. 3, Jun. 2008.

[11] J. Li, T. Wang, and Y. Zhang, "Face Detection using SURF Cascad", *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011.

[12] J. Ribas and S. Lei, "Rate control in dct video coding for low-delay communications", *IEEE Transations on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 172-185, Feb. 1999.