# A NOVEL WEIGHT-BASED URQ SCHEME FOR PERCEPTUAL VIDEO CODING OF CONVERSATIONAL VIDEO IN HEVC

*Shengxi Li, Mai Xu, Xin Deng, and Zulin Wang*

School of Electronic and Information Engineering, Beihang University
Beijing, 100191, China
ShengxiLi2014@gmail.com, MaiXu@buaa.edu.cn, cindydeng1991@gmail.com, wzulin@buaa.edu.cn

## ABSTRACT

In this paper, we propose a novel weight-based unified rate-quantization (URQ) scheme for rate control in state-of-the-art HEVC standard, to improve its perceived visual quality for conversational videos. In conventional rate control of HEVC, a pixel-wise URQ scheme is proposed by introducing the concept of bit per pixel (bpp). This scheme is able to assign different amounts of bits to the blocks with various sizes, thus well suitable for flexible picture partition of HEVC. However, bpp does not reflect the visual importance of each pixel. Therefore, we propose a novel weight-based URQ scheme to take into account the visual importance for rate control in HEVC. In combination with the weight map acquired from a novel hierarchical perceptual model of face, such a scheme is capable of allocating more bits to the face and much more bits to the facial features, by using bit per weight (bpw) instead of bpp. As a result, the visual quality of face, especially facial features, can be improved such that perceptual video coding is achieved for HEVC. Finally, the experimental results validate such improvement.

***Index Terms*—** HEVC, rate control, pixel-wise URQ scheme, perceptual video coding

## 1. INTRODUCTION

Most recently, high efficiency video coding (HEVC) standard [1], has been formally established, to relieve the bandwidth-hungry issue. With flexible picture partition, parallel coding and some other cutting-edge technologies, HEVC has eminent compression performance, much better than the preceding H.264/AVC standard [2].

Rate control is a crucial module in HEVC. More specifically, at a given bit-rate, the visual quality should be optimized via reasonably allocating the bits to various blocks and frames. There are many rate control algorithms for different video coding standards (e.g. TM5 for MPEG-2, VM8 for MPEG-4 and JVT-N046 for H.264). For HEVC, a pixel-wise rate control scheme was developed in [3], by using a pixel-wise unified rate-quantization (URQ) model. Since the proposed scheme works at pixel level, it can be easily applied to blocks with different sizes, thus reducing the encoder complexity.

The URQ model for rate control in HEVC assumes that each pixel has equal visual importance during bit allocation. In fact, when a person watches a generic video, he/she may not pay attention to the whole scene according to human visual system (HVS) [4]. In other words, a small region around a point of fixation, called region-of-interest (ROI) region, is concerned most (e.g. [5], [6]), while the peripheral region is captured at low resolution. In light of this phenomenon, a large amount of bits can be saved by reducing the bit number in non-ROI regions, meanwhile minimizing the distortion in ROI regions. Consequently, the perceptual redundancy can be further exploited to improve perceived visual quality through optimizing bit allocation.

There has been a growing interest in rate control in perceptual video coding [7] [8] [9] [10] [11] for previous video coding standards. In H.263, a perceptual rate control (PRC) scheme [8] was proposed. In this scheme, a perceptual sensitive weight map of the conversational scene is yielded to assign more bits to ROI regions via reducing the value of quantization parameters (QPs). Afterwards, for H.264/AVC, a novel source allocation method [10] was proposed to optimize the subjective rate-distortion-complexity performance of conversational video coding, by improving the visual quality of face region extracted by the skin-tone algorithm [12]. However, to our best knowledge, there is no perceptual video coding approach for rate control in state-of-the-art HEVC standard.

In this paper, we propose a novel weight-based rate control scheme on the basis of URQ model, in order to improve perceived visual quality for conversational video compressed by HEVC. To be more specific, similar to [10], we consider face in conversational video as ROI region. Since it is intuitive that facial features (e.g. nose, mouth and eyes) are more important than other regions within a face while their distortion is larger than other regions, we further impose more importance on facial features in HEVC. We thus develop a hierarchical perception (HP) model of face, which can yield a pixel-wise weight map to indicate the unequal importance of

This work was supported by NSFC under grant number 61202139 and China 973 program under grant number 2013CB329006.

background, non-facial-features[1] and facial features. Then, the core of our weight-based URQ scheme is that given the weight map by HP model, bit per weight (bpw), instead of bit per pixel (bpp) in conventional URQ scheme, is utilized to calculate QP of each block. This way, more bits can be allocated to the face region, especially facial features. Finally, the perceived visual quality of conversational video compressed by HEVC can be improved, with the visual quality of facial features and non-facial-features being enhanced at different levels according to the pixel-wise weight map.

## 2. OVERVIEW OF THE PIXEL-WISE URQ SCHEME

The key issue of rate control in video coding is computing QPs[2], which aims at minimizing the distortion of a compressed video at a given bit-rate. However, there exists a chicken and egg dilemma between the actually generated bits and the QPs. In order to solve such a dilemma, a quadratic pixel-wise URQ model [3] has been proposed to calculate QPs, based on the predicted target bits and image complexity before actual encoding. It is important to recognize that such a rate control scheme is developed at pixel level, and as such, it can be easily applied at frame and unit (similar to block in H.264) levels.

At unit level, the calculation of QPs, which determines the actual bit allocation in the URQ scheme, is related to target bit $T_j$ of the $j$th largest coding unit (LCU), and its predicted mean absolute difference (MAD) $\mathrm{MAD}_{\mathrm{pred,j}}$. Assume that the number of pixels in the $j$th LCU is $M$. Its QP value $\mathrm{QP_j}$ can be acquired by solving the following quadratic equation:

$$\frac{T_j}{M} = a \times \frac{\mathrm{MAD}_{\mathrm{pred,j}}}{\mathrm{QP_j}} + b \times \frac{\mathrm{MAD}_{\mathrm{pred,j}}}{\mathrm{QP_j^2}}, \quad (1)$$

where $a$ and $b$ are the first-order and second-order parameters of URQ model, which may be updated by a linear regression method after encoding each frame [13]. $\mathrm{MAD}_{\mathrm{pred,j}}$ represents the image complexity, which is determined once the current frame is settled. Thus, the left task to estimate $\mathrm{QP_j}$ is calculating $T_j$. In the URQ scheme, $T_j$ is determined by two factors: target bits based on remaining bits $\hat{T}_j$ and target bits based on buffer status $\tilde{T}_j$. Thus, $T_j$ can be formulated:

$$T_j = \beta \times \hat{T}_j + (1-\beta) \times \tilde{T}_j, \quad (2)$$

where $\beta$ $(= 0.5)$ represents the tradeoff between the buffer status and remaining bits on computing $T_j$. In this paper, we focus on $\hat{T}_j$, and the details about obtaining $\tilde{T}_j$ are presented in [3]. Before acquiring $\hat{T}_j$, bpp is introduced as a key term in the URQ scheme. For the $j$th LCU, bpp is denoted as $\mathrm{bpp_j}$:

$$\mathrm{bpp_j} = \frac{B_j}{N_j}. \quad (3)$$

Note that $\mathrm{bpp_j}$ is updated before the encoding of the $j$th LCU. In addition, $N_j$ represents the number of un-encoded pixels in the frame, and $B_j$ denotes the target bits for the $j$th and its subsequent LCUs in the frame. $B_j$ is initialized to be the whole target bits of the frame when encoding the first LCU, and it is updated in the following way:

$$B_j = B_{j-1} - A_{j-1}, \quad (4)$$

where $A_{j-1}$ is the actual bits to encode the $(j-1)$th LCU.

When $\mathrm{bpp_j}$ is updated by (3) and (4), $\hat{T}_j$ can be calculated with

$$\hat{T}_j = \mathrm{bpp_j} \times M. \quad (5)$$

Then, to avoid sudden change of QPs, the URQ scheme adds the following boundary to smooth QP values:

$$\mathrm{QP_j} = \max\{\mathrm{QP_{avg,j}} - 2, \min\{\mathrm{QP_{avg,j}} + 2, \mathrm{QP_j}\}\}, \quad (6)$$

where $\mathrm{QP_{avg,j}}$ is the average QP value of the neighboring (already encoded) LCUs of the $j$th LCU. At last, rate control can be achieved in the URQ scheme via the output of QPs. In addition, from the definition of bpp, each pixel is endowed the equal importance. However, it is intuitive that the pixels in the ROI (face and facial features) and non-ROI (background) regions have the unequal importance during bit allocation. Thus, the URQ scheme wastes many bits on encoding the non-ROI regions, to which human pay less attention. In Section 3, we develop a HP model, to identify the importance of each pixel. On the basis of HP model, the weight-based URQ scheme is proposed in Section 4.

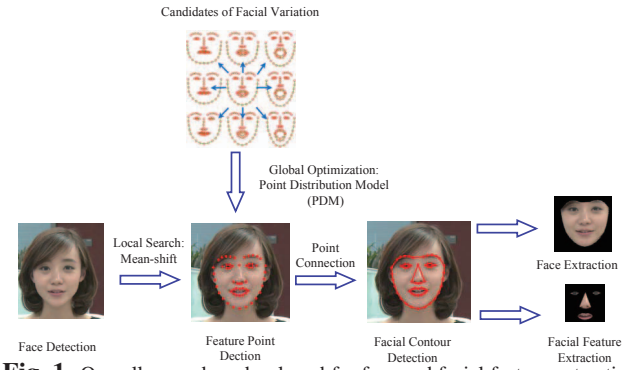



**Fig. 1**. Overall procedure developed for face and facial feature extraction.

## 3. HIERARCHICAL PERCEPTION MODEL

In this section, we mainly focus on the HP model, to provide the pixel-wise weight map for our weight-based URQ scheme. Since the perceptual mechanism of HVS has a long way to go yet, the precise identification of ROI regions, indeed, is intractable in perceptual video coding. Fortunately, we have an important cue for the perception model in conversational video coding, i.e. considering the face as the ROI regions. It is intuitive that facial features are more important than other regions within a face. Therefore, the goal of using HP model is to stress on the further decomposition of ROI regions, in order to arrange different importance weights inside ROI regions. Specifically, for better correlation with HVS, a conversational video frame needs to be decomposed into background and face. A face region can be further decomposed into several sub-regions and even several sub-sub-regions (if necessary) with different importance.

[1] Non-facial-features mean the regions within face excluding the facial features.

[2] A small value of QP yields the small quantization step (QS), which means that more bits can be generated for high quality video.

In this paper, face is further decomposed into facial features and non-facial-features. Towards such a decomposition, the face and its facial features can be extracted using the procedure in Figure 1. As shown in this figure, after face detection [14], several key feature points are located in the video by combining the local detector and global optimization together. Benefiting from the most recent success on computer vision, we employ a real-time face alignment method [15] to track these feature points in the face, and an example of the detection results can be seen in Figure 1. Then, the contours of the face and its facial features are achieved via connecting the related feature points. Finally, the regions of the face and the facial features can be extracted upon these contours.

Next, based on the extracted contours, the HP model is developed, as illustrated in Figure 2. It can be seen from this figure that the whole region of conversational video is decomposed hierarchically into several subregions, with three layers. In this hierarchy, the face and background are separated at the second layer. The face is further decomposed into several facial features and non-facial-features at the third layer. Besides, each node in Figure 2 is associated with a weight for its importance[3]. Then, in HP model, each pixel in a video frame falls into one leaf node, and the importance weight of a pixel is computed by summing up the weights of its leaf node and all the corresponding root nodes. For example, if a pixel belongs to nose, the weight of its leaf node is 0 and the weights of its root nodes (i.e. face at layer 2 and whole region at layer 1) are both 1. Therefore, the weight of this pixel is 2. This way, the pixel-wise weight map can be produced using the HP model upon the extracted face and facial features. We assume that the weights are $\{w_n\}_{n=1}^{N}$ in a weight map for a video frame with $N$ pixels.

Finally, according to HVS [11], the pixel-wise weight map can be refined via introducing Gaussian model (GM) to the weights of pixels around eye fixation point (i.e. regions with large weights). We define $\Delta d_n$ as the distance of $n$th pixel to the edge of the nearest facial feature (but not falling into it). Assume that $w^i$ is the weight[4] of the node in the HP model for the facial feature that is closest to the $n$th pixel; $\sigma^i$ is the standard deviation for the decay of $w^i$ around contour of the facial feature. Then, the weights of pixels around facial features can be updated with GM by adding the following Gaussian increment:

$$\Delta w_n = w^i e^{-\frac{1}{2}(\frac{\Delta d_n}{\sigma^i})^2}, \tag{7}$$

[3]The values of weights in each node are determined by the visual importance in HVS and the actual distortion performance of video sequences compressed by HEVC. On the one hand, HVS points out that the face region, especially mouth and eyes, attracts more human attention, which is validated by our eye tracking experiment (displayed in web-site http://www.ee.buaa.edu.cn/xumfiles/). This is also pointed out by [16]. On the other hand, the rate-distortion performance of regions of face, especially eyes and mouth, is inferior. As such, more weights (=3) are imposed in these regions.

[4]In Figure 2, $i = 1$ represents eyes node: $w^1 = 3$; $i = 2$ represents mouth node: $w^2 = 3$; $i = 3$ represents nose node: $w^3 = 0$.

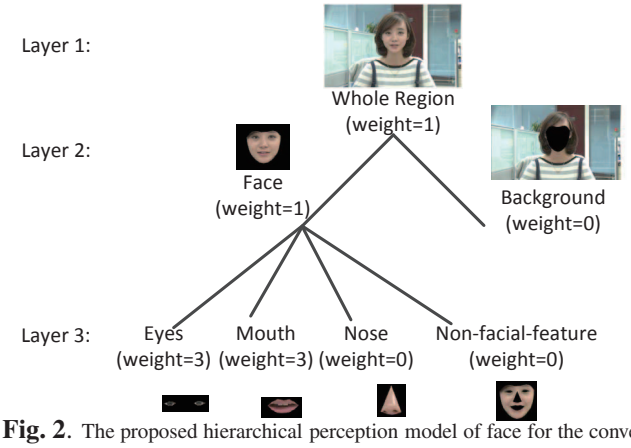



**Fig. 2**. The proposed hierarchical perception model of face for the conversational video.


**Fig. 3**. An example of the pixel-wise weight map for the video frame in Figure 2.

into their original weights. Note that only weights of the pixels around eyes and mouth are updated. After GM refining, the pixel-wise weight map can be output and then used for the weight-based URQ scheme discussed in Section 4. Figure 3 shows an example of the pixel-wise map refined by GM.

## 4. THE PROPOSED WEIGHT-BASED URQ SCHEME

In this section, a weight-based URQ scheme is proposed, by introducing a new term bpw, instead of bpp, to allocate bits according to the HP model of Section 3. First, let us look at the calculation of bpw, which is initialized by

$$\mathrm{bpw} = \frac{\hat{T}}{\sum_{n=1}^{N} w_n}, \tag{8}$$

where $\hat{T}$ is the target bit for the current frame and it can be estimated by [3] before encoding this frame. In addition, $w_n$ stands for the weight of the $n$th pixel, and $N$ represents the number of pixels in the frame. After obtaining the value of bpw, we separate the target bits of whole frame into the target bit for ROI (face and facial features) regions and non-ROI (background) regions, defined by $\hat{T}'$ and $\hat{T}''$, respectively:

$$\begin{cases} \hat{T}' + \hat{T}'' = \hat{T} \\ \hat{T}'' = \frac{\sum_{n\in\mathbf{n}''} w_n \mathrm{bpw}}{c \times \sum_{n\in\mathbf{n}'} w_n \mathrm{bpw}} \hat{T}', \end{cases} \tag{9}$$

where $\mathbf{n}''$ denotes the indices of background pixels, the weights of which are equivalent to 1; $\mathbf{n}'$ means indices of the facial pixels, the weights of which are greater than 1. In (9), $c$ is the parameter balancing the bits assigned to ROI and

non-ROI regions. By solving (9), we can obtain:

$$\hat{T}' = \frac{c \times \sum_{n \in \mathbf{n}'} w_n}{c \times \sum_{n \in \mathbf{n}'} w_n + \sum_{n \in \mathbf{n}''} w_n} \hat{T}, \tag{10}$$

$$\hat{T}'' = \frac{\sum_{n \in \mathbf{n}''} w_n}{c \times \sum_{n \in \mathbf{n}'} w_n + \sum_{n \in \mathbf{n}''} w_n} \hat{T}. \tag{11}$$

For LCUs of non-ROI regions, our weight-based scheme is reduced to conventional pixel-wise URQ scheme, since the weights of pixels in the background are equivalent to 1. For LCUs of ROI regions, the following step is applied for bit allocation, given target bits $\hat{T}'$. Before encoding these LCUs, $\mathrm{bpw_j}$ for the $j$th LCU needs to be updated:

$$\mathrm{bpw_j} = \frac{B'_j}{\sum_{n \in \mathbf{n}'^j} w_n}, \tag{12}$$

where $B'_j$ and $\mathbf{n}'^j$ define the target bits and pixel indices, respectively, for the $j$th and its subsequent LCUs of ROI regions. Note that $B'_j$ is initialized to be $\hat{T}'$, when encoding the first LCU of ROI regions. Afterwards, $B'_j$ needs to be updated for the following LCUs of ROI regions, with the way same as calculating $B_j$ in (4). Then, $\hat{T}'_j$, target bits of the $j$th LCU in the ROI regions, is computed:

$$\hat{T}'_j = \sum_{n \in \mathbf{n}_j} w_n \times \mathrm{bpw_j}, \tag{13}$$

where $\mathbf{n}_j$ is the pixel indices in the $j$th LCU. As seen from (13), the LCU with large bpw and $w_n$ is assigned more target bits. Therefore, ROI regions, especially the facial features, are emphasized with more bits.

Finally, for the $j$th LCU in ROI regions, similar to (2), total target bits $T'_j$, can be estimated by $\hat{T}'_j + \tilde{T}'_j$, where $\tilde{T}'_j$ is the target bits regarding the buffer status. For the $j$th LCU in non-ROI regions, total target bits $T''_j$ can be computed using the same way as Section 2. Then, the rate control at unit level can be achieved in our weight-based URQ scheme, by outputting QPs for all LCUs with (1).

Note that our scheme iterates over the steps of estimating QPs and updating target bits, for encoding each LCU in scan order. The main difference between our scheme and pixel-wise URQ scheme [3] is the step of updating target bits, in which our scheme introduces a novel term bpw, instead of the conventional bpp, catering for the HP model. Then, by adjusting target bits to modify QP values, the pixel importance can be taken into account such that our scheme adaptively increases the number of bits in the ROI regions, in accordance with HVS.

In addition, in order to avoid the sudden change of QP values to alleviate block effect, we set boundary for QPs of each LCU. However, different from (6), the QP boundary in our scheme is also influenced by importance of each LCU,

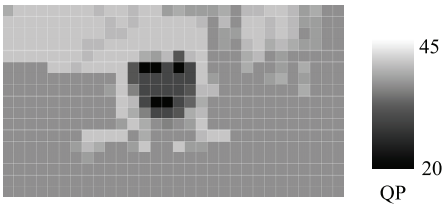



**Fig. 4**. Example of QPs at LCU level for the video frame of Figure 2, output by our weight-based URQ scheme. They are obtained using the weight map of Figure 3. Each block stands for a $64 \times 64$ block, except the blocks of the last row. The intensities of each block indicate the QP values.

and it is rewritten as:

$$\mathrm{QP_j} = \max\{\mathrm{QP_{avg,j}} - \frac{2}{M} \sum_{n \in \mathbf{n}_j} w_n,$$

$$\min\{\mathrm{QP_{avg,j}} + \frac{2}{M} \sum_{n \in \mathbf{n}_j} w_n, \mathrm{QP_j}\}\}, \tag{14}$$

which is adaptive to the average weight of each LCU. Since the average weight of each LCU in ROI regions is greater than that of non-ROI regions, its QP boundary is broader than that in non-ROI regions as observed in (14), further improving the visual quality in ROI regions. Similarly, visual quality improvement in the facial features (that have larger average weights) may be greater than non-facial-features.

One example for the map of QPs by our scheme is demonstrated in Figure 4. We can observe from this figure that eyes and mouth have the smallest QPs, followed by other facial regions, and the background has the largest QPs. Such QP output shows that our weight-based URQ scheme assigns more bits to ROI regions, satisfying the HP model.

## 5. EXPERIMENTAL RESULTS

In this section, experimental results are presented to validate the proposed weight-based URQ scheme. We have used four test video sequences, two CIF conversational video sequences: *Akiyo* and *Foreman*, and two high definition (HD) conversational video sequences: *Yan* and *Simo*. For the test HD videos, we captured two raw conversational video sequences at $1920 \times 1080$ with 150 frames, using a Sony XDCAM-PDW-700 camera, since there is no standard HD conversational video available. After video capture, we converted all video frames into the color images with RGB components (8 bits per component) in BMP format. Then, all the BMP images were assembled into a video sequence in YUV format with 4:2:0 sampling. The video sequences are freely downloadable[5].

In addition, we used the HEVC test model (HM) 9.0 software [17] with its default pixel-wise URQ scheme [3] as the reference approach. Our weight-based URQ scheme was implemented on HM 9.0 for comparison. The parameter settings of HM 9.0 software are listed in Table 1. Furthermore, the first frame in each GOP is regarded as I-frame and the other frames are treated as P-frames. Besides, in our approach, we

[5]Our web-site is http://www.ee.buaa.edu.cn/xumfiles/.

**Table 1**. The parameters for video coding

| | |
|---|---|
| Total number of frames | 150 frames |
| Frame rate | 25 fps |
| GOP size | 4 frame |
| LCU size | $64 \times 64$ pixels |
| Maximum LCU depth | 4 |
| Search Range for ME | 64 pixels |
| SAO | Enabled |

empirically set $(\sigma^i)^2 = \sqrt{m^i}$ for refining the weight map in (7), where $m^i$ is the number of pixels belonging to the $i$th facial feature. Moreover, parameter $c$ in (9) was appropriately tuned to 5.

Figure 5 shows the rate-distortion performance of the conventional URQ and our schemes in face, background and whole regions. As can be easily seen from this figure, our scheme outperforms the conventional approach in terms of average Y-PSNR of the face regions, for all the video sequences at different bit-rates. As the cost, the average Y-PSNR of the background is decreased in our approach. However, thanks to HVS, human beings mostly concentrate on the face (i.e. ROI region), while little eyesight stays in the background (i.e. non-ROI region) long. Consequently, the increase of distortion in background cannot set off a huge tempest for the integral subjective video quality perception.

Moreover, we plot Figure 6 to show the further improvement within ROI regions, that is, non-facial-features and facial features (e.g. nose, mouth and eyes). It can be observed from this figure that the rate-distortion performance of facial features is significantly improved at various bit-rates, for both CIF and HD videos, in our approach. Furthermore, the distortion difference within a face for the HD video sequences is alleviated so that the overall perceived visual quality can be further refined as it agrees with the HVS.

In addition, we compare the conventional URQ and our schemes in terms of subjective quality. Figure 7 demonstrates the 103th reconstructed frames of *Foreman* compressed at 40 kbps, and Figure 8 shows the 42th reconstructed frames of *Yan* compressed at 100 kbps. As expected, our approach is capable of yielding more favorable visual quality in the face, especially in the facial features, with the sharper edges and less blurred texture.

In summary, the experimental results clearly demonstrate the effectiveness of our approach in improving perceived visual quality, comparing with the conventional URQ scheme on HM 9.0.

## 6. CONCLUSIONS

In this paper, we have proposed a novel weight-based URQ scheme for conversational video in HEVC, to improve its perceived visual quality. First, we have established a HP model to reflect the importance of visual content for conversational video. On the basis of the HP model, a novel weight-based URQ scheme was proposed, by using $\mathrm{bpw}$ rather than $\mathrm{bpp}$ to take into account the visual importance of each pixel. Thus, in accord with HVS, the perceived visual quality is improved by our approach, as more bits are assigned to ROI regions, especially the facial features. Finally, we have conducted several experiments which validate the superior rate-distortion performance of our weight-based URQ scheme on the HEVC platform.

## 7. REFERENCES


[1] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22(12), pp. 1649 – 1668, Dec. 2012.

[2] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13(7), pp. 560 – 576, Jul. 2003.

[3] H. Choi, J. Yoo, J. Nam, D. Sim, and I. V. Bajic, "Pixel-wise unified rate-quantization model for multi-level rate control," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7(6), pp. 1112 – 1123, Dec. 2013.

[4] C. Blakemore and F. W. Campbell, "On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images," *The Journal of physiology*, vol. 203, no. 1, pp. 237–260, 1969.

[5] W. S. Geisler and J. S. Perry, "A real-time foveated multi-resolution system for low-bandwidth video communication," in *Proceedings of the SPIE: The International Society for Optical Engineering*, 1998, vol. 3299, pp. 294–305.

[6] Z. Wang and A. C Bovik, "Embedded foveation image coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 10, pp. 1397–1410, 2001.

[7] D. Chai and K. N. Ngan, "Face segmentation using skin-color map in videophone applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9(4), pp. 551 – 564, Jun. 1999.

[8] X. Yang, W. Lin, Z. Lu, X. Lin, S. Rahardja, E. Ong, and S. Yao, "Rate control for videophone using local perceptual cues," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15(4), pp. 496 – 507, Apr. 2005.

[9] C.-W. Tang, "Spatiotemporal visual considerations for video coding," *IEEE Transactions on Multimedia*, vol. 9(2), pp. 231 – 238, Apr. 2007.

[10] Y. Liu, Z. G. Li, and Y. C. Soh, "Region-of-interest based resource allocation for conversational video communication of H.264/AVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18(1), pp. 134 – 139, Jan. 2008.

[11] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image and Vision Computing*, vol. 29(1), pp. 1 – 14, Jan 2011.

[12] R.-L. Hsu, M. Abdel-Mottaleb, and A.K. Jain, "Face detection in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24(5), pp. 696 – 706, May 2002.

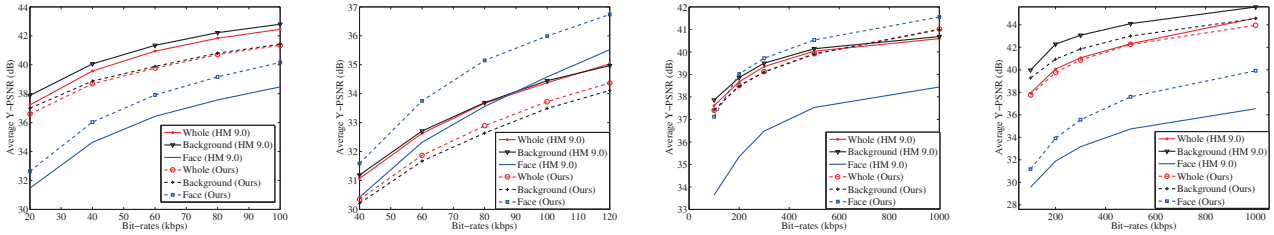
[13] Y. Liu, Z. G. Li, and Y. C. Soh, "A novel rate control scheme for low delay video communication of H.264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17(1), pp. 68 – 78, Jan 2007.

[14] S. Z. Li and A. K. Jain, *Handbook of Face Recogntition*, Springer, 2011.

[15] J. M. Saragihand, S. S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proceeding of ICCV*, 2009, pp. 1034–1041.

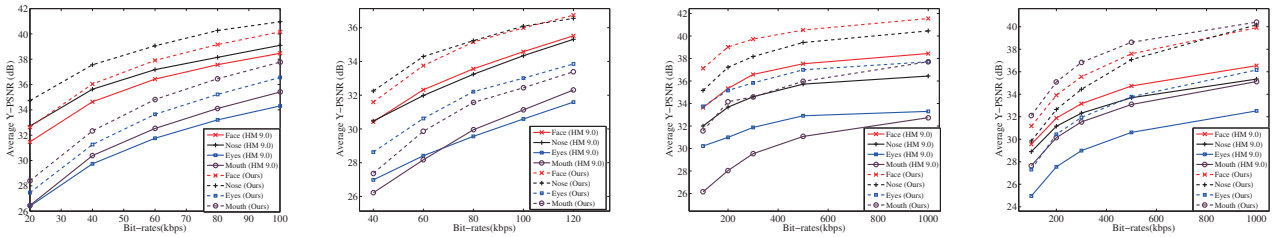
[16] S. W. Janik, A. R. Wellens, M. L. Goldberg, and L. F. Dell'Osso, "Eyes as the center of focus in the visual examination of human faces," *Perceptual and Motor Skills*, vol. 47, no. 3, pp. 857–858, 1978.

[17] JCT-VC, "HM 9.0," http://hevc.hhi.fraunhofer.de/.



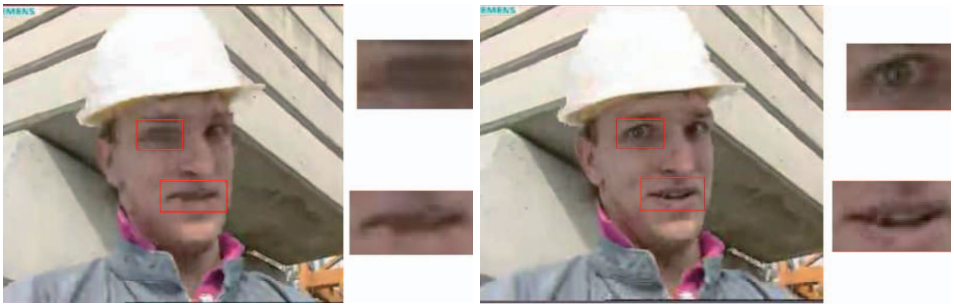

(a) *Akiyo* (CIF) (b) *Foreman* (CIF) (c) *Yan* (1080p HD) (d) *Simo* (1080p CIF)

**Fig. 5**. Rate-distortion performance comparison for face, background and whole regions, between the conventional HM 9.0 approach and our approach on compressing four conversational video sequences.



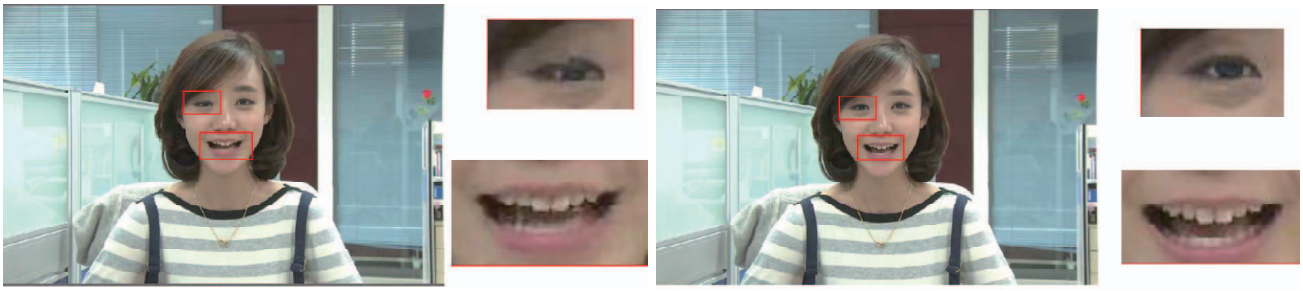

(a) *Akiyo* (CIF) (b) *Foreman* (CIF) (c) *Yan* (1080p HD) (d) *Simo* (1080p CIF)

**Fig. 6**. Rate-distortion performance comparison for the regions of face and facial features, between the conventional HM 9.0 approach and our approach on compressing four conversational video sequences.



(a) HM9.0 (b) Our approach

**Fig. 7**. Visual quality comparison of Foreman (CIF resolution). (a) and (b) show its 103th decoded frames compressed at 40 kbps by HM 9.0 and our approach, respectively. In (a), the average Y-PSNRs of the whole region, face, mouth, eyes and nose in HM 9.0 are 31.07 dB, 30.42 dB, 26.22 dB, 26.98 dB and 30.47 dB. In (b), the average Y-PSNRs of the whole region, face, mouth, eyes and nose in our approach are 30.35 dB, 31.59 dB, 27.37 dB, 28.63 dB and 32.25 dB.



(a) HM9.0 (b) Our approach

**Fig. 8**. Visual quality comparison of Yan (1080p HD resolution). (a) and (b) show its 42th decoded frames compressed at 200 kbps by HM 9.0 and our approach, respectively. In (a), the average Y-PSNRs of the whole region, face, mouth, eyes and nose in HM 9.0 are 38.17 dB, 35.34 dB, 28.04 dB, 30.99 dB and 33.68 dB. In (b), the average Y-PSNRs of the whole region, face, mouth, eyes and nose in our approach are 38.50 dB, 39.02 dB, 34.16 dB, 35.16 dB and 37.22 dB.