# Region-of-Interest Based Conversational HEVC Coding with Hierarchical Perception Model of Face

Mai Xu, *Member, IEEE*, Xin Deng, *Student Member, IEEE*, Shengxi Li, *Student Member, IEEE*, and Zulin Wang, *Member, IEEE*

***Abstract*—**

**fi**

**fi**

***Index Terms*—**

## I. Introduction

OWADAYS, plenty of conversational video products, such as FaceTime, are flooding into our lives, facilitating the visual communications for humans. On the other hand, the past decade has witnessed a great evolution of ever-increasing video resolutions and screen display sizes. Accordingly, the conversational videos, in particular at high resolutions, are causing the bandwidth bottleneck. Fortunately, High Efficiency Video Coding (HEVC) standard [1], also called H.265, has been formally established, to provide higher compression efficiency for supporting such bandwidth-hungry applications. With the flexible picture partition, parallel coding and some other cutting-edge techniques, HEVC has eminent compression performance, much better than the preceding H.264/AVC standard [2].

Nevertheless, high resolution video delivery, especially to meet low bit-rates of mobile devices, still poses the great challenging problem of compression efficiency for HEVC. In fact, there still remains much perceptual redundancy in HEVC, since human attentions do not focus on the whole scene, but only a small region of fixation called region-of-interest (ROI) region. For example, it has been found out in [3] that humans normally perceive clearly a small region of 2–5° of the visual angle. Thereby, perceptual video coding [4] provides an efficient solution towards lower bit-rate video coding, which keeps acceptable distortion in ROI regions, but at the expense of some visual quality degradation in non-ROI regions.

Recently, there has been a growing interest in perceptual video coding [5]–[7]. More specifically, Lee and Bovik [5] proposed to use an eye tracker to obtain the fixation points as ROI regions, for the earlier H.263 standard. However, it is impractical to have the eye tracker available during the video encoding process. Automatic ROI region extraction, based on the perception model of human visual system (HVS), is thus the key issue for perceptual video coding. Intuitively, the important cue for the perception model in conversational video coding is extracting faces as ROI regions. Then, a perceptual rate control scheme [6] was proposed to reduce the quantization parameter (QP) values of skin regions in H.263, with a block-wise sensitive weight map of the conversational scene. Afterwards, a novel resource allocation method [7] was proposed for H.264/AVC standard to optimize the subjective rate-distortion-complexity performance of conversational video coding, by improving the visual quality of facial regions

Manuscript received September 14, 2013; revised January 20, 2014; accepted March 25, 2014. Date of publication April 02, 2014; date of current version May 13, 2014. This work was supported by the National Science Foundation of China (NSFC) under Grant 61202139 and the China 973 program under Grant 2013CB329006. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Karen Egiazarian.

The authors are with the School of Electronic and Information Engineering, Beihang University, Beijing, 100191 China (e-mail: maixu@buaa.edu.cn; cindydeng1991@gmail.com; shengxili2014@gmail.com; wzulin@buaa.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2014.2314864

■ 288), and they are not specifically designed for high resolution videos. In general, the existing perceptual approaches on video coding are neither suitable for the HEVC standard nor the video services with high resolutions.

In this paper, we propose a novel perceptual approach for conversational HEVC coding, in order to improve its perceived visual quality, especially at high resolutions. On one hand, as seen from Fig. 1, the number of pixels representing each facial feature, such as eyes and mouth, in an HD video is comparable with

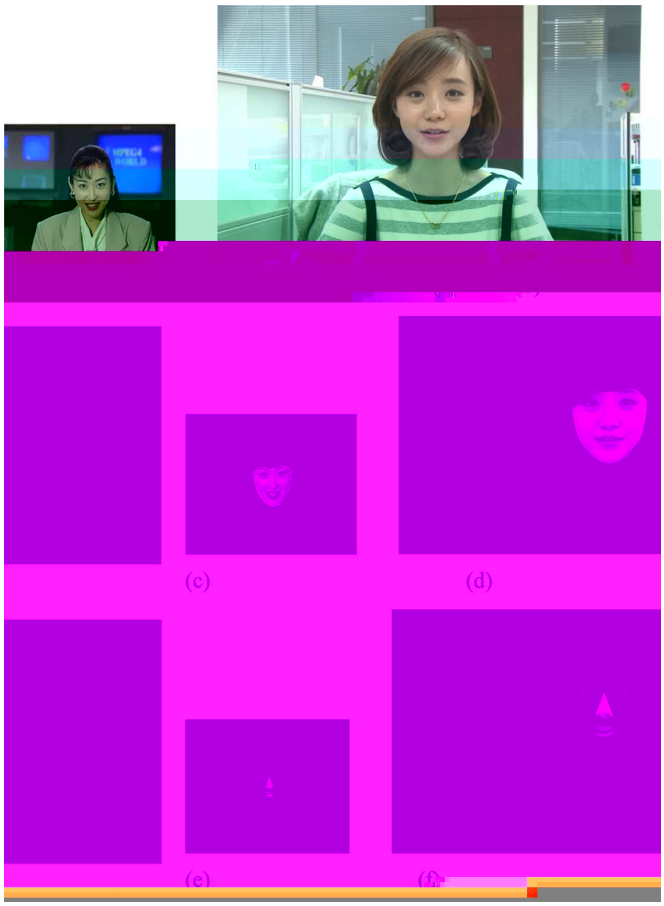



Fig. 1. (a) and (b): The original conversational video sequences of *Akiyo* and *Yan* (only showing one random frame) with the resolutions being $352 \times 288$ and $1920 \times 1080$, respectively. (c): The face (4810 pixels) detected from (a). (d): The face (99896 pixels) obtained from (b). (e): The eyes (58 pixels), nose (248 pixels) and mouth (241 pixels) extracted from (a). (f): The eyes (1828 pixels), nose (5186 pixels) and mouth (4508 pixels) extracted from (b). Note that the face and facial features are extracted using the method of Section III-B. (a) *Akiyo* (CIF), (b) *Yan* (1080p HD), (c) Extracted face of CIF video, (d) Extracted face of HD video, (e) Extracted facial features of CIF video, (f) Extracted facial features of HD video.

that representing the whole face in a CIF video. On the other hand, it is intuitive that the facial features are the noticeable regions in a face to attract human attentions. It is thus desirable that the visual quality of facial features is superior to other ROI regions in conversational HEVC coding. However, we find out that the rate-distortion performance of the regions of facial features, compressed by HEVC, is even worse than other regions. To overcome such a drawback, in addition to the enhancement of facial quality, the proposed approach in this paper further improves the visual quality of facial features for conversational HEVC coding.

The basic philosophy of this paper is twofold: (1) The pixel-wise weight maps of conversational video are yielded, constrained by a hierarchical perception model of face (HP model); (2) coding tree unit (CTU) structure and QPs are then adaptively adjusted on the basis of the pixel-wise weight maps, to allow the unequal importance during video coding. The main contributions of this paper are listed in the following.

- We analyze the rate-distortion performance of conversational HEVC coding in the regions of background, face, and facial features, for illustrating the necessity of our work. Then, we propose an HP model, which indicates the unequal importance of facial features, non-facial-features, and background[1] with a pixel-wise weight map. Towards such a perception model, an extraction method of face and facial features is presented in light of the face alignment algorithm [9].
- We develop an adaptive CTU partition structure to reduce encoding complexity of conversational HEVC coding. Intrinsically, the CTU partition structure introduced by HEVC improves the rate-distortion performance, but at the cost of computational complexity. In our approach, given the pixel-wise weight maps, the CTUs of facial regions have to be partitioned in details through setting large maximum depths for largest coding units (LCU), to maintain the perceived visual quality. Then, rough partitions are applied to other CTUs with small maximum LCU depths, thus reducing a great deal of encoding complexity in HEVC.
- We propose a rate control scheme using the pixel-wise weight maps upon our HP model. Generally speaking, the core of our scheme is to allocate more bits to ROI regions, i.e., face and facial features, by utilizing the weight-based unified rate-quantization (URQ) scheme, instead of the conventional pixel-based URQ scheme. Beyond, the perceived visual quality of conversational HEVC coding can be improved, with the visual quality of ROI regions enhanced at different levels according to the pixel-wise weight maps.

The outline of this paper is given as follows. In Section II, we briefly review the previous work on perceptual video coding. In Section III, the details of the proposed HP model are discussed. Based on Section III, Section IV develops an adaptive CTU partition structure for HEVC, which is capable of decreasing its encoding complexity. Afterwards, Section V proposes a weight-based URQ scheme to improve the visual quality of ROI regions in conversational HEVC coding. Finally, Section VI shows some experimental results and Section VII concludes this paper.

## II. Previous Work

For several years, there has been a great deal of interest in perceptual video coding [4]. Generally speaking, the main idea of perceptual video coding is increasing the coding efficiency via removing the perceptual redundancy. That is, it imposes high priority on ROI regions while allowing more distortion in non-ROI regions. Therefore, the perceptual video coding involves two major parts: perception model with ROI regions and video coding implementation upon the ROI regions.

For the perception model, many methods have emerged in perceptual video coding. At the beginning, human-machine interaction methods [5], [10]–[12] were adopted to obtain ROI regions for the perception model of video coding. For example, in 1990s, Kortum and Geisler [10] developed a real-time multi-resolution system, which utilizes an eye tracker to record the

[1]Non-facial-features are defined as the regions of face excluding the facial features. Background, as non-facial-regions, is defined as the regions of whole scene except the face.

foveation points of a human observer on the receiver and then applies a corresponding foveation filter in video coding of the sender. Later, several advanced approaches on using the eye tracker for video coding have been proposed [5], [11], [12]. However, it is hard to implement such approaches due to the fact that an eye-tracking apparatus is normally unavailable at the receiver. From the perspective of psychology, many approaches on the perception model for video coding [13]–[16] have been proposed to predict which regions in a video can attract human attentions, according to HVS mechanisms. A representative work is the saliency-based attention prediction [15], in light of Itti-Koch attention model [17], for detecting ROI regions in perceptual video coding. Such an approach produces a guidance map to locate ROI regions with the top salient values in a video, and on the basis of the guidance map, a bit allocation scheme varies the QP value of each block for video coding. Since the study on HVS is still in progress, it is rather difficult to incorporate the HVS mechanisms into perceptual video coding. Evidently, human face is an important cue [18] for perceptual video coding, especially under conversational scenarios. Thereby, many approaches [6], [7], [19]–[21] define faces as the ROI regions for conversational video coding. Actually, this kind of perception model is very effective in conversational video applications, as it benefits from the recent success of face detection [22] in computer vision community. However, the above perception model does not consider the facial features as more important regions, and these facial features in HD conversational videos usually take up a great number of pixels. In order to achieve better perceived visual quality, these facial features, in particular for HD videos, need to be endowed more importance.

For video coding implementation, the previous approaches were developed in dichotomy: either pre-processing or embedded encoding. Preprocessing approaches [10], [16], [23]–[25] are straightforward, as they directly reduce the unimportant information of input video by applying a non-uniform distortion filter in a scene. For example, [16] divides a scene into foreground and background, and then the background, as the non-ROI region, is blurred with a filter (i.e., more distortion is imposed) to save some bits for video coding. Besides, imitating human vision, a foveation filter was applied in [23] to increase the blurring effect along with the distance between the considered pixel and the eye fixation point. Obviously, the advantage of the preprocessing approaches is that they are independent of the existing video coding standards and they are thus easily applied. However, these approaches introduce the blurring effects in non-ROI regions, which produce the obvious degradation of visual quality in those regions. Embedded encoding approaches [6], [7], [19], [26], [27] have been developed to increase the bit allocation in ROI regions by reducing their corresponding QP values, thereby improving the perceived visual quality of video coding. For the earlier H.261 standard, two quantizers were developed in [19]. The facial regions, as ROI regions, are allocated more target bits through adjusting these two quantizers. For H.263, a perceptual rate control scheme [6] was proposed, in combination with the local perceptual cues, to improve the visual quality of skin regions in a conversational video. For H.264/AVC, a novel source allocation method [7] was proposed to enhance the subjective rate-distortion-complexity performance of conversational video coding. However, to our best knowledge, there is no perceptual encoding implementation for the latest HEVC standard [1].

In this paper, we propose an approach on perceptual video coding, embedded on the state-of-the-art HEVC standard. Such an approach is based on a novel HP model, and it is capable of improving the visual quality of ROI regions (including facial features and non-facial-features) to obtain better perceived visual quality, with even less encoding complexity.

## III. Hierarchical Perception Model of Face

The study on perceptual mechanisms of HVS has a long way to go yet. It is still intractable to precisely identify the ROI regions for perceptual video coding. Fortunately, we have an important cue for the perception model in conversational video coding, i.e., considering the face as ROI regions. First, we investigate in Section III-A the rate-distortion performance of different regions in conversational HEVC coding. With the investigation results, we argue that the HP model is indeed necessary. In Section III-B, we propose an HP model, which is in accord with the perceptual mechanisms of HVS. At last, pixel-wise weight maps are generated to impose the unequal importance on each pixel in a conversational video.

### *A. The Necessity of Hierarchical Perception Model of Face*

Admittedly, HEVC, as the latest video coding standard, has improved rate-distortion performance a lot in comparison with the preceding H.264/AVC. However, it still has some undesirable defects. In fact, HEVC adopts a similar coding structure as H.264/AVC. This results in a challenge resembling H.264/AVC, that is the ill-suited bit-allocation problem which may lead to unsatisfactory visual perception performance. It is therefore worth investigating the visual quality of different regions in conversational HEVC coding.

Here, we tested the rate-distortion characteristic of the conversational HEVC coding on four conversational video sequences[2]: *Akiyo* (CIF), *Foreman* (CIF), *Simo* (1080P HD), and *Yan* (1080P HD). We applied HM 9.0 software [28] of HEVC to compress each video sequence, using the default pixel-based URQ scheme [29] for rate control. The parameter settings are to be presented in Section VI-A. In addition, the algorithm for automatic extraction of face and facial features is to be presented in Section III-B. Examples of the extraction results can be seen in Fig. 1.

Now, we examine the changes of Y-PSNRs under a range of bit-rates to investigate the rate-distortion performance of HEVC. Since the regions of face, especially facial features, usually attract most of human attentions, they are the key regions for investigation. The average Y-PSNRs of the whole region, background, face, and facial features at various bit-rates are plotted in Fig. 2. As can be seen there, it is obvious that the average Y-PSNRs of face are smaller than those of the whole region and background for nearly all video sequences at different bit-rates, except the *Foreman* sequence at high bit-rates. This result reveals the necessity of the previous work on the

[2]Since there is no standard HD conversational video sequence available, we captured four raw HD video sequences using the method described in Section VI-A.

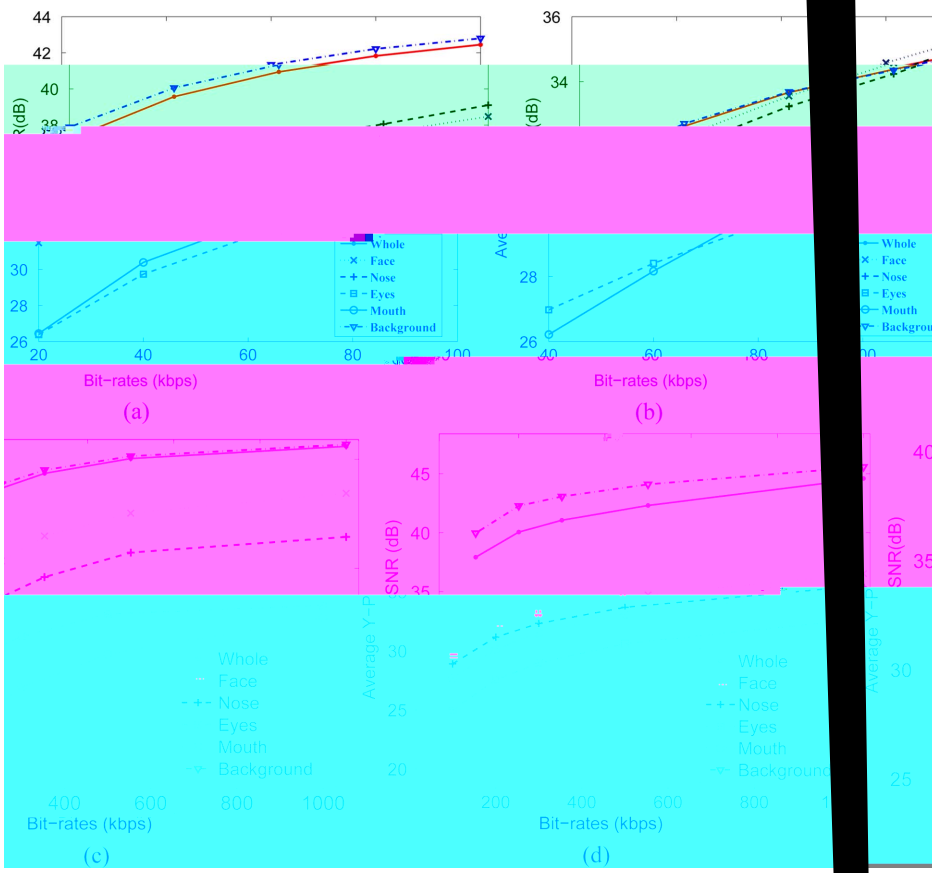



Fig. 2. The rate-distortion performance of the whole region, background, face, nose, eyes and mouth for four convers nal vide its default pixel-based URQ scheme for rate control. (a) *Akiyo* (CIF), (b) *Foreman* (CIF), (c) *Yan* (1080p HD), (d) *Sim* 080p H

perceptual video coding [6], [7]. It can be further observed from Fig. 2 that the average Y-PSNRs of the eyes and mouth are much lower than the face/whole region. Such results trigger our work on the HP model, which further splits the facial region into several subregions (e.g., the eyes and mouth) and then assigns unequal importance weights to them.

### *B. The Proposed Hierarchical Perception Model of Face*

From now on, we mainly focus on the HP model. As aforementioned, it is necessary to further decompose face into several facial features and non-facial-features. Note that non-facial-features are defined as the regions of face excluding the facial features. Towards such a decomposition, the face and its facial features can be extracted using the procedure of Fig. 3. As shown in this figure, after face detection [22], several key feature points are located in the video by combining the local detection and global optimization together. Next, the contours of the face and its facial features are achieved via connecting the related feature points. Finally, the regions of the face and facial features are extracted upon their contours. Indeed, the detection of feature points is a key issue for extracting the regions of face and facial features. Benefiting from the most recent success on computer vision, we employ a real-time face alignment method [9] to track the feature points in face. An example of the detection results can be seen in Fig. 3. In the following, we briefly review the work of the face alignment method [9].

For global optimization, the point distribution model (PDM) of key feature points needs to be coined before tracking these points in a video. Assume th $\{\mathbf{p}_t\}_{t=}^{T}$ dinates of each key feature po t. Ther rameterize the deformable fa on the $\{\bar{\mathbf{p}}_t\}_{t=1}^{T}$ by

$$\mathbf{p}_t = s\mathbf{R}(\bar{\mathbf{p}} + \mathbf{\Phi}_t\mathbf{q}$$

with the parameters of scale rotati a set of non-rigid parameters In (1 candidates of non-rigid facial riatio as the *prior*, we can select the y fea favor the locally detected fea e poi using the least-square fit:canb

$$\min_{\mathbf{p}_t} \sum_{=1}^{T} \|\mathbf{p}_t$$

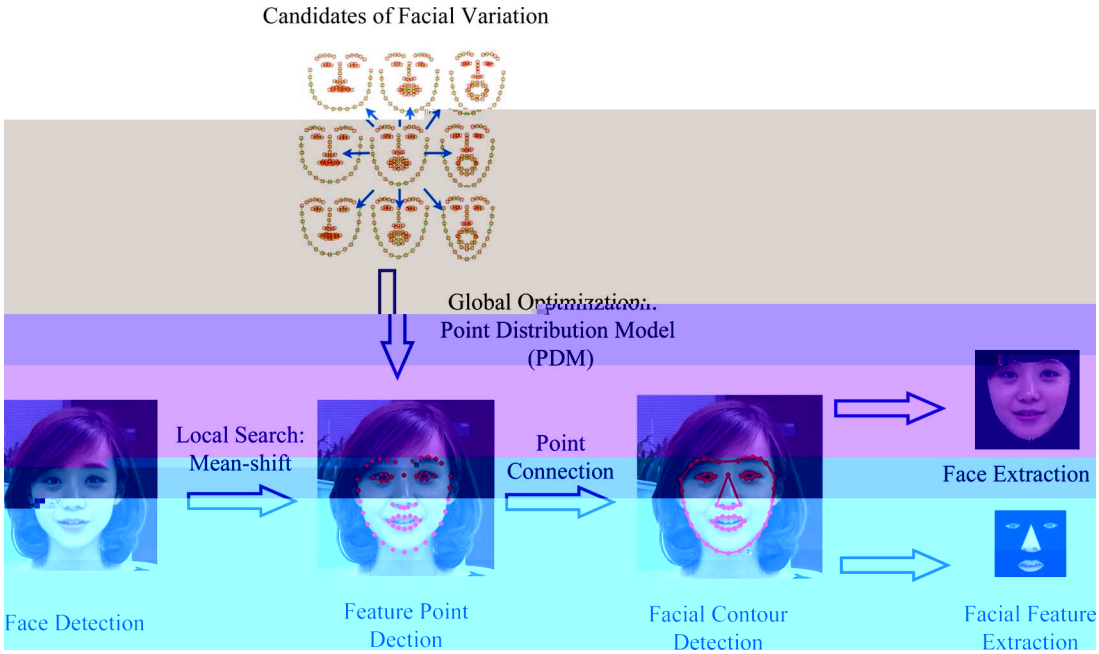



Fig. 3. Overall procedure developed for face and facial feature extraction.

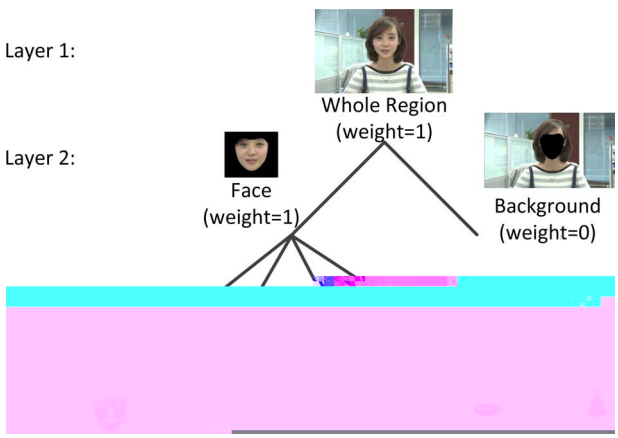



Fig. 4. The proposed hierarchical perception model of face for the conversational video.

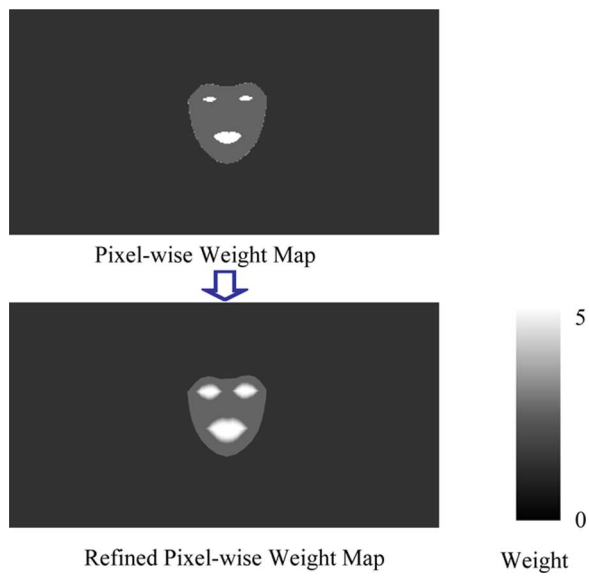



Fig. 5. The pixel-wise weight map for Fig. 1-(b).

66-point PDM is applied in our approach to extract face and facial features.

Next, based on the extracted face and facial features, the HP model is developed, as illustrated in Fig. 4. It can be seen from this figure that the whole region of each conversational video frame is decomposed hierarchically into several subregions, belonging to three perceptual layers. In this hierarchy, the background and face are separated at the second layer. Different from the conventional approaches [6], [7], the face is further decomposed into several facial features and non-facial-features at the third layer. Moreover, each node in Fig. 4 is associated with a weight for its importance. The values of importance weights of nodes are determined by HVS and the rate-distortion performance of different regions compressed by HEVC. To be more specific, the study of HVS [31] has shown that the eyes and mouth, in particular eyes, attract much more eye fixation points than nose and non-facial-features, when humans look at a conversational scene. Besides, we also used a Tobii T60 eye tracker to identify eye fixation points over several conversional video clips, and we recorded the eye fixation points of 12 observers over 18 conversational video clips, with 30 seconds per video clip. From the recorded results, we can see that nose and non-facial-features draw similar amount of human attentions, much less than eyes and mouth[3]. However, the visual quality of eyes and mouth is inferior in comparison with nose and non-facial-features, as presented in the above subsection. Therefore, zero weight is set for nose and non-falcial-features, and larger weight $(= 3)$ is assigned to eyes and mouth.

Then, HP model is worked out to obtain the pixel-wise weight map, which indicates the varying importance of different regions in a conversational scene. In the HP model, each pixel in a video frame falls into one leaf node, and the importance weight

[3]We have put the detailed eye-tracking results online: http://www.ee.buaa.edu.cn/xumfiles.

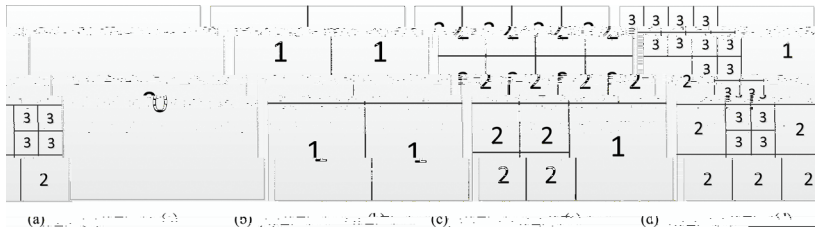


Fig. 6. Example of CTU partition structure, which divides an LCU (size: $64 \times 64$) into several CUs with different sizes. Note that in (a), (b), (c), and (d) each block indicates a CU and the number inside a block stands for its depth. (a) Maximum depth: 0/1/2/3, (b) Maximum depth: 1/2/3, (c) Maximum depth: 2/3, (d) Maximum depth: 3.

of a pixel can be 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 the pixel belonging to nose 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. Here, one example of the pixel-wise weight map is shown in Fig. 5 (the upper one).

Finally, according to HVS [15], 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 $v_i$ is the weight of the node for the $i$th facial feature[4] in the HP model; $\sigma_i$ is the standard deviation for the decay of $v_i$ around contour of the $i$th facial feature. Then, the weights of pixels around each facial feature can be updated with GM, by adding the following Gaussian increment:

$$\Delta w_n = v_i e^{-\frac{1}{2}(\frac{\Delta d_n}{\sigma_i})^2}, \tag{3}$$

into their original weights. Note that only weights of the pixels around eyes and mouth are updated, due to their corresponding $v_i > 0$. After GM refining, the pixel-wise weight map can be output and then used for the ROI-based video coding discussed in the following sections. Fig. 5 (the bottom one) shows an example of the pixel-wise map refined by GM.

## IV. ROI-Based Adaptive CTU Partition Structure for HEVC

In this section, we present a novel ROI-based adaptive CTU partition structure for HEVC, based on the HP model above. In Section IV-A, we first review the conventional CTU partition structure employed in HEVC, as the foundation of the proposed adaptive CTU partition structure. Then, in Section IV-B, we provide the detailed information about the ROI-based adaptive CTU partition structure.

### A. The CTU Partition Structure in HEVC

One of the most significant contributions in HEVC is the CTU partition structure. It has been pointed out [32] that the size of $16 \times 16$ macroblocks in H.264/AVC standard is too monotonous

[4]In Fig. 4, $i = 1$ represents eyes node: $v_1 = 3$; $i = 2$ represents mouth node: $v_2 = 3$; $i = 3$ represents nose node: $v_3 = 0$.

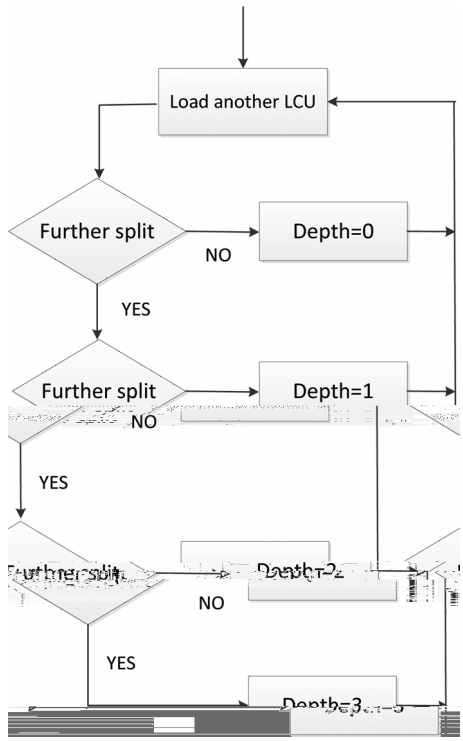



Fig. 7. The procedure of subdividing an LCU into CUs with different depths. Note that the "further split" is conducted on the CUs from the last splitting.

to adapt to video content at different resolutions, which may contain large smooth areas or small specific details. By contrast, the CTU partition structure of HEVC can offer more flexible block sizes, ranging from $64 \times 64$ to $8 \times 8$, thus suitable for both smooth and detailed regions. To be more specific, there are four splitting depths in HEVC, i.e., from 0 to 3, for dividing an LCU into several coding units (CUs) at different levels. In each LCU, four equally sized CUs may be recursively partitioned with different depths. Then, each CU can be used as the basic unit for both intra-coding and inter-coding. An example of splitting an LCU into CUs with different sizes is shown in Fig. 6. From this figure, it can be seen that even though the LCU is allowed to be divided into small CUs given the same maximum depth, not all of them can go there. The determination process of LCU splitting is shown in Fig. 7. From this figure, it can be seen that an LCU can be divided into CUs at different depths. The condition for further splitting is that the rate-distortion cost of the current CU is larger than the sum of the cost of its four split CUs. Only the CUs satisfying such a condition can be further split to reach a larger depth. Note that the maximum depth of an LCU may be achieved after several iterations of further splitting.

The CTU partition structure is capable of improving the rate-distortion performance of HEVC, owing to the flexible block partition, as discussed above. However, it consumes an enormous amount of computational time when splitting each LCU into CUs, because of the computation on the rate-distortion cost of each possible CU. Fortunately, HEVC offers the optional setting of the maximum LCU depth, to which the depths of all CUs cannot exceed. This may significantly reduce encoding complexity. For instance, only the partitions of Fig. 6(a) and (b) are allowed once the maximum LCU depth is chosen to be 1. Then, only the first two splitting steps of Fig. 7 need to be conducted in CTU partition structure, thus gaining a great deal of encoding time.

In fact, most of the time, the maximum LCU depth does not need to be very large, especially for non-ROI regions. According to the HVS, the detailed information is not necessary in non-ROI regions. Whether the depth is large or small for CUs in non-ROI regions, therefore, has little effect on the whole perceived visual quality, while small depth is able to save a lot of computational time. We may make constraint on the maximum LCU depths in non-ROI regions to reduce the encoding complexity. We discuss the specific algorithm by proposing our ROI-based adaptive CTU partition structure in the next subsection.

### *B. The Proposed ROI-Based Adaptive CTU Partition Structure*

As aforementioned, one feasible way to reduce the encoding complexity of conversational HEVC coding is assigning different maximum depths to the LCUs of different regions, according to their relative importance. In other words, the less important the LCU is, the smaller depth it is assigned with. It has been previously discussed that the pixel with larger weight implies that it is relatively more important. Here, we define by $\lambda_j$ the average weight of the $j$th LCU, and it can be calculated on the basis of the weight map of Section III-B using

$$\lambda_j = \frac{1}{M} \sum_{n \in \mathbf{n}_j} w_n, \tag{4}$$

where $\mathbf{n}_j$ means the set of pixel indices in the $j$th LCU of a video frame, and $M$ is the number of pixels in the LCU.

After calculating the values of $\{\lambda_j\}_{j=1}^{J}$ for all $J$ LCUs of a video frame, their maximum depths $\{z_j\}_{j=1}^{J}$ can be obtained with the following equation:

$$z_j = \begin{cases} 1 & \text{if } \lambda_j \leq \theta_1 \\ 2 & \text{if } \theta_1 < \lambda_j \leq \theta_2 \\ 3 & \text{if } \lambda_j > \theta_2. \end{cases}$$

where $a$ and $b$ are the first-order and second-order parameters of URQ model, which can be updated by a linear regression method [33] after encoding each frame. Recall that $M$ is the number of pixels in each LCU. Then, after solving (6), $\mathrm{QP_j}$ can be obtained as

$$\mathrm{QP_j} = \frac{a \cdot \mathrm{MAD_{pred,j}} + \sqrt{a^2 \cdot \mathrm{MAD_{pred,j}}^2 + 4b \cdot \mathrm{MAD_{pred,j}} \cdot \frac{T_j}{M}}}{\frac{2T_j}{M}}. \tag{7}$$

Then, the only task left for estimating $\mathrm{QP_j}$ is to determine target bits $T_j$ for each LCU. $T_j$ is related to two factors: the buffer status and the actual remaining bits, with the following equation:

$$T_j = \beta \cdot \hat{T}_j + (1 - \beta) \cdot \tilde{T}_j, \tag{8}$$

where $\hat{T}$ [redacted] for each

calculating $B_j$ in Section V-A. Different from the bpp in (9), bpw is able to incorporate the weights of pixels given the HP model.

Then, the target bits for each facial LCU, with respect to the remaining bits, can be written as [illegible]

$$\hat{T}'_j = \sum$$

[illegible] (18) where regions of facial features. Finally, similar [illegible] LCU of pixels. Clearly, the facial features can be adaptive

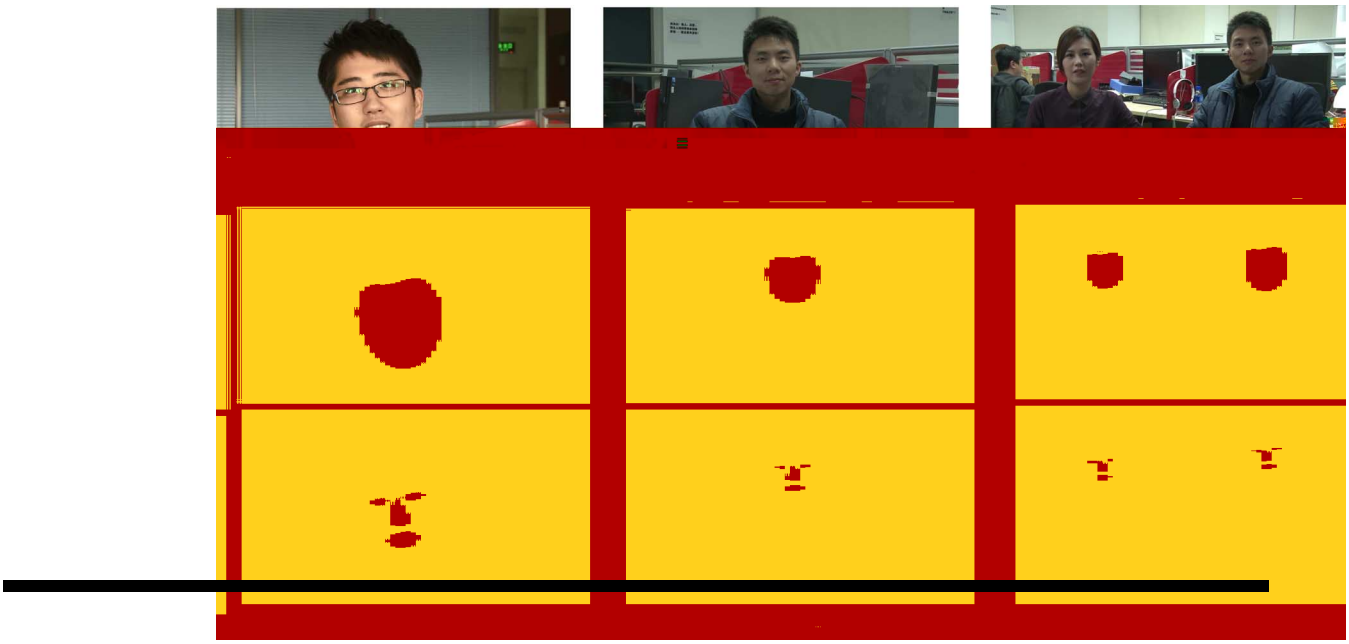


Fig. 10. First row: Original conversational video sequences of *Simo*, *Lee* and *Couple* (only showing one random frame) with the resolution 1920 ■ 1080. Second row: Extracted faces of the conversational video sequences of the first row. Third row: Extracted facial features of the conversational video sequences of the first row. (a) *Simo*, (b) *Lee*, (c) *Couple*.

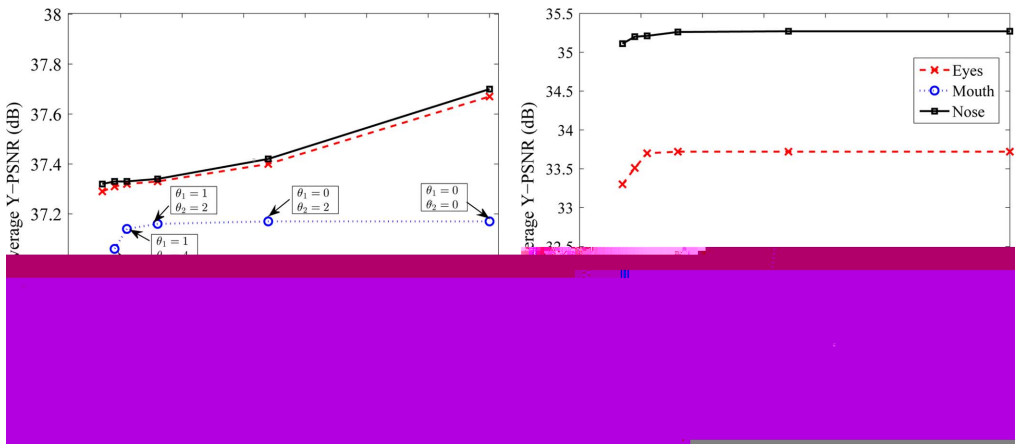



Fig. 11. Complexity-distortion curves of compressing video sequence *Yan* with our approach at 100 kbps. (a) Whole, background, and facial regions, (b) Facial features.

TABLE II
ENCODING TIME REDUCTION OF OUR APPROACH OVER CONVENTIONAL HM 9.0 APPROACH ON CIF VIDEO SEQUENCES

| Video Sequences | *Akiyo* | | | | | *Foreman* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Bit-rates (kbps) | 20 | 40 | 60 | 80 | 100 | 40 | 60 | 80 | 100 | 120 |
| Encoding time reduction (%) | 19.4 | 22.1 | 23.8 | 20.0 | 20.6 | 22.2 | 22.1 | 21.5 | 21.8 | 22.9 |
| Average Y-PSNR improvement in face (dB) | 1.15 | 1.32 | 1.43 | 1.59 | 1.61 | 1.13 | 1.43 | 1.51 | 1.38 | 1.16 |

TABLE III
ENCODING TIME REDUCTION OF OUR APPROACH OVER CONVENTIONAL HM 9.0 APPROACH ON HD VIDEO SEQUENCES

| Video Sequences | *Yan* | | | | | *Simo* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Bit-rates (kbps) | 100 | 200 | 300 | 500 | 1000 | 100 | 200 | 300 | 500 | 1000 |
| Encoding time reduction (%) | 54.0 | 54.5 | 53.4 | 51.8 | 53.0 | 58.7 | 57.1 | 56.0 | 54.8 | 53.8 |

that the values in the horizontal axis are normalized encoding time. From this figure, we can see that along with the increasing encoding time, the average Y-PSNR of the whole region of decoded video enhances slightly. However, the quality enhancement in the regions of face/facial features almost stops once the normalized time arrives at 0.46, in which the corresponding settings of $\theta_1$ and $\theta_2$ are 1 and 2, respectively. This phenomenon is possibly due to the fact that the maximum LCU depths of facial regions almost reach the largest value (i.e., 3) in such settings. Therefore, in the subsequent experiments, we set thresholds $\theta_1$ and $\theta$

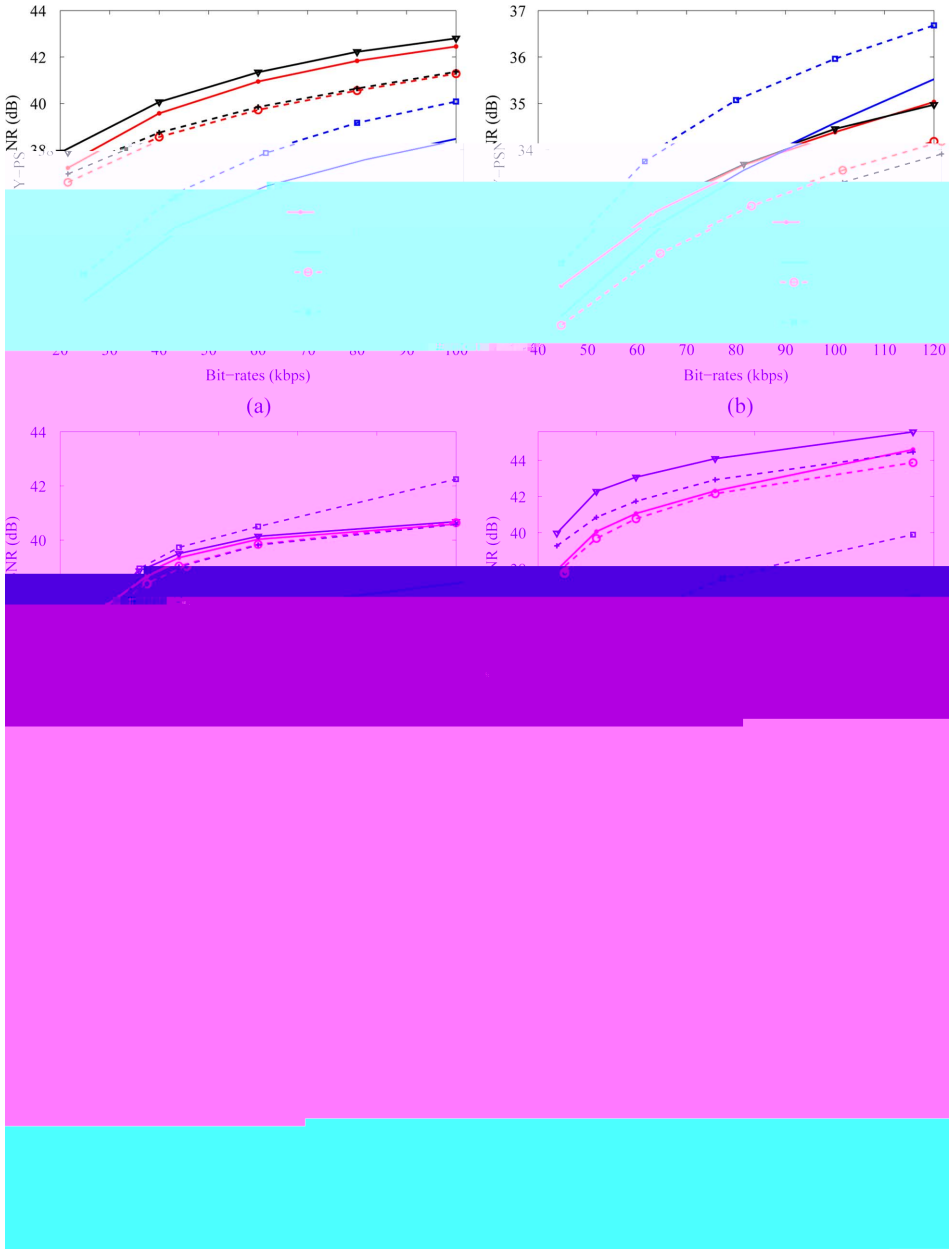



Fig. 12. Rate-distortion performance comparison for face, background and whole regions, between the conventional HM 9.0 and our approaches, on compressing six conversational video sequences. (a) *Akiyo* (CIF), (b) *Foreman* (CIF), (c) *Yan* (1080p HD), (d) *Simo* (1080p HD), (e) *Lee* (1080p HD), (f) *Couple* (1080p HD).

on all the video sequences compressed by our and the conventional HM 9.0 approaches at different bit-rates.

Here, we adopted a single stimulus continuous quality scale (SSCQS) procedure, proposed by Rec. ITU-R BT.500 [34], to rate the subjective quality. The experiment we conducted was divided into two sessions. The first session included only CIF test video sequences: *Akiyo* and *Foreman*, while the second one was comprised of HD test sequences: *Yan*, *Simo*, *Lee*, and *Couple*. Note that the uncompressed and compressed video sequences in each session were displayed in a random order. Before each session, the observers were required to view 5 other training videos (one training video per quality scale) to help them better understand the subjective quality assessment. 12 observers (4 females and 8 males), aging from 20 to 45, were involved in this test. We used a 23″ DELL U2312HM LCD monitor with its resolution being $1920 \times 1080$ to display the videos. The viewing distance was set to be approximately three times of the video height for rational evaluation. The quality rate scales for observers to evaluate after viewing are: excellent (100-81), good (80-61), fair (60-41), poor (40-21), and bad (20-1).

After the subjective evaluation, we computed Difference Mean Opinion Scores (DMOS), indicating the visual difference between the compressed and uncompressed videos. The smaller the value of DMOS is, the better subjective quality the compressed video sequence has. Then, Table IV compares the average DMOS values of all compressed video sequences. From this table, we can see that the DMOS values of our approach are rather smaller than those of the conventional HM 9.0 approach. In other words, our approach can provide higher subjective video quality, especially for HD videos at relatively low bit-rates.

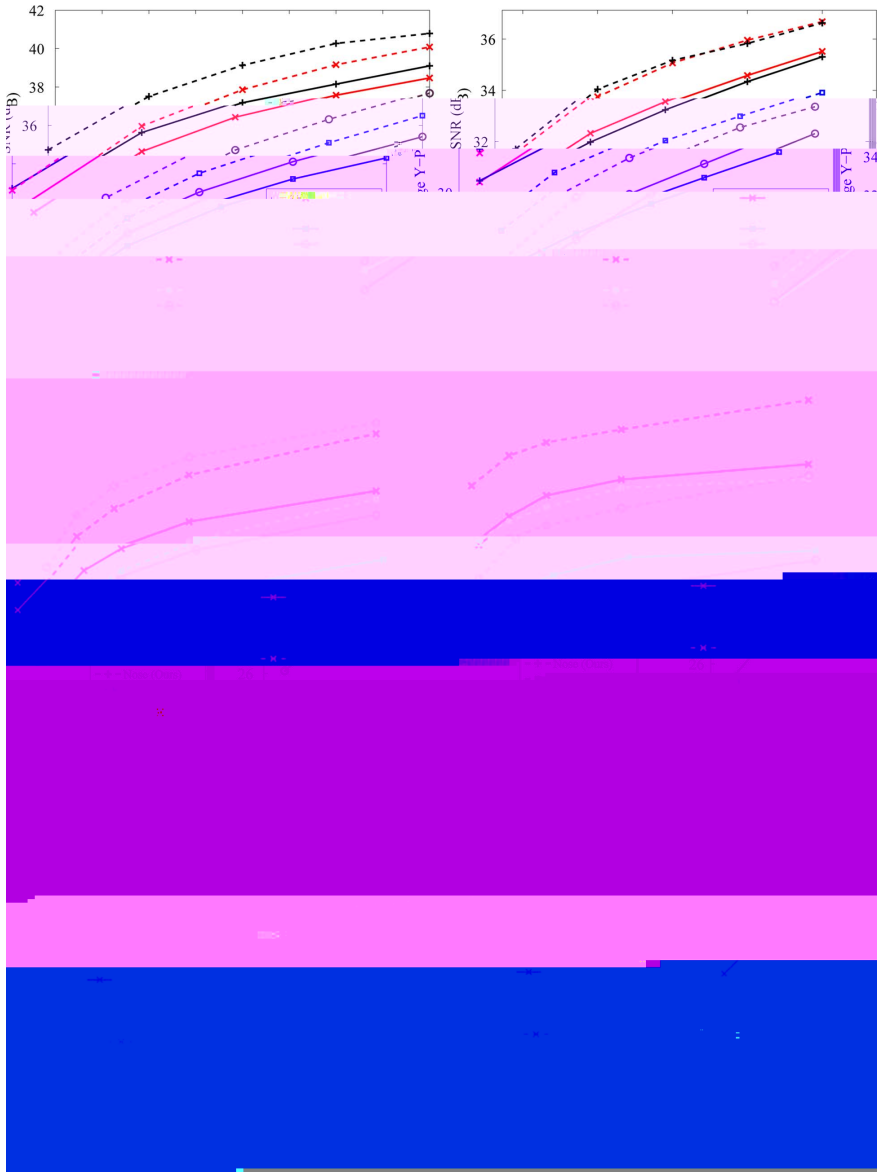



Fig. 13. Rate-distortion performance comparison for the regions of face and facial features, between the conventional HM 9.0 and our approaches, on compressing six conversational video sequences. (a) *Akiyo* (CIF), (b) *Foreman* (CIF), (c) *Yan* (1080p HD), (d) *Simo* (1080p HD), (e) *Lee* (1080p HD), (f) *Couple* (1080p HD).


Fig. 14. Visual quality comparison of *Foreman* (CIF resolution). (a) and (b) show its 110th decoded frames compressed at 40 kbps by our and HM 9.0 approaches, respectively. In (a), the average Y-PSNRs of the background, face, mouth, eyes and nose in HM 9.0 are 31.19 dB, 30.42 dB, 26.22 dB, 26.98 dB and 30.47 dB. In (b), the average Y-PSNRs of the background, face, mouth, eyes and nose in our approach are 30.07 dB, 31.55 dB, 27.34 dB, 28.52 dB and 31.71 dB. (a) HM 9.0, (b) Our approach.

In summary, our subjective results here, together with the previous objective results reported in Sections VI-B and VI-C, illustrate that our approach on conversational HEVC coding performs better in terms of both encoding time and perceived visual quality.

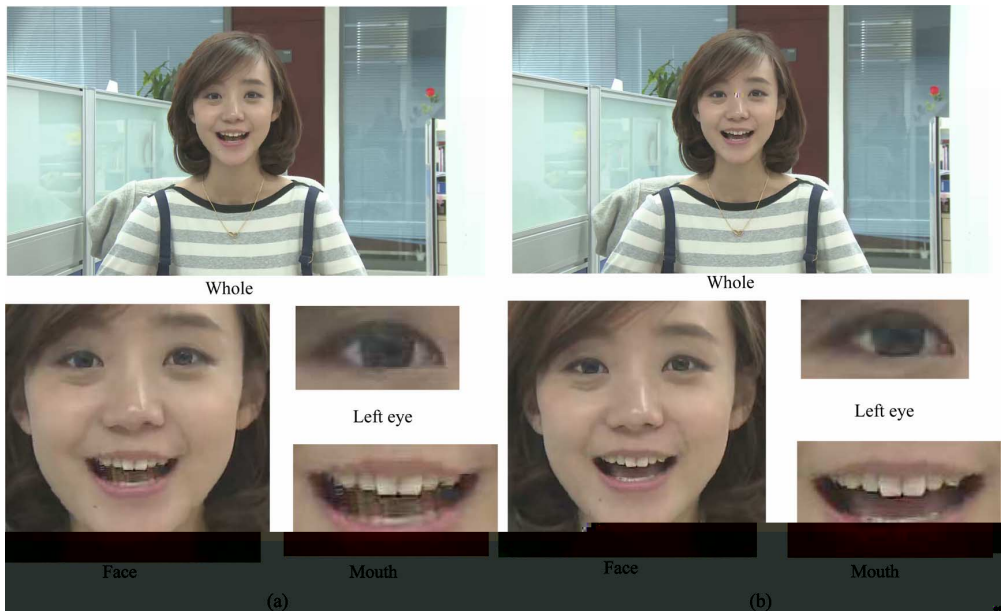



Fig. 15. Visual quality comparison of *Yan* (1080p HD resolution). (a) and (b) show its 20th decoded frames compressed at 100 kbps by our and HM 9.0 approaches, respectively. In (a), the average Y-PSNRs of the background, face, mouth, eyes and nose in HM 9.0 are 37.86 dB, 33.64 dB, 26.15 dB, 29.22 dB and 32.00 dB. In (b), the average Y-PSNRs of the background, face, mouth, eyes and nose in our approach are 37.33 dB, 37.16 dB, 31.65 dB, 33.72 dB and 35.26 dB. (a) HM 9.0, (b) Our approach.

TABLE IV
DMOS COMPARISON OF OUR AND CONVENTIONAL HM 9.0 APPROACHES

| Sequences | Resolution | Bit-rates (kbps) | HM 9.0 DMOS | Our DMOS | DMOS Difference |
|---|---|---|---|---|---|
| *Akiyo* | 352×288 | 20 | 59.06 | 50.16 | -8.90 |
| | | 40 | 34.66 | 23.45 | -11.21 |
| *Foreman* | 352×288 | 60 | 73.44 | 62.93 | -10.51 |
| | | 80 | 57.62 | 43.71 | -13.91 |
| *Yan* | 1920×1080 | 100 | 71.88 | 46.15 | -25.73 |
| | | 300 | 46.22 | 31.46 | -14.76 |
| *Simo* | 1920×1080 | 100 | 71.63 | 57.73 | -13.90 |
| | | 300 | 54.35 | 38.16 | -16.19 |
| *Lee* | 1920×1080 | 100 | 67.23 | 40.62 | -26.61 |
| | | 300 | 45.41 | 29.15 | -16.26 |
| *Couple* | 1920×1080 | 100 | 73.78 | 46.41 | -27.37 |
| | | 300 | 47.39 | 28.16 | -19.23 |

## VII. CONCLUSION

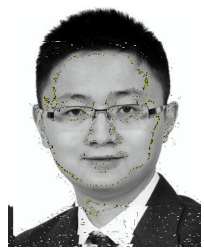
In this paper, we have proposed an ROI-based perceptual video coding approach, for improving the perceived visual quality of conversational videos on the HEVC platform. It was argued that in conversational HEVC coding, the rate-distortion performance of some important subregions inside an ROI region (i.e., mouth and eyes in a face) is inferior to other ROI subregions. Therefore, in contrast with the previous perceptual video coding approaches, our approach endows the unequal importance within the facial region to emphasize its facial features, by proposing a perception model called HP model. Benefiting from the HP model, an ROI-based adaptive CTU partition structure was developed to reduce the encoding complexity of HEVC, while maintaining the visual quality in the facial region, especially in the regions of facial features. Furthermore, we proposed a weight-based URQ scheme, instead of the previous pixel-based URQ scheme in HEVC, to adaptively assign bits according to the HP model. This way, the visual quality of face and facial features, in conversational HEVC coding, is enhanced to varying degrees in accordance with their importance weights, thereby greatly improving the overall perceived visual quality. Finally, the experimental results demonstrated that our approach considerably outperforms the conventional HEVC approach, in terms of both encoding time and perceived visual quality, for conversational video coding.

Our work in its present form merely focuses on the rate control scheme at LCU level. Therefore, it is hard to significantly improve the visual quality of facial features for the videos at low resolutions, due to the fact that the sizes of facial features may be even smaller than the sizes of LCUs. On the other hand, the rate control at CU level, in keeping with the flexible CTU partition structure of HEVC, provides a promising trend for the future work.

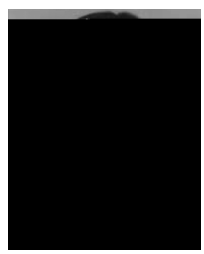
## REFERENCES

[1] G. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
[2] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[3] B. Wandell, *Foundations of Vision*. Sunderland, MA, USA: Sinauer, 1995.
[4] J. Lee and T. Ebrahimi, "Perceptual video compression: A survey," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 684–697, Oct. 2012.
[5] S. Lee and A. C. Bovik, "Fast algorithms for foveated video processing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 2, pp. 149–161, Feb. 2003.
[6] X. Yang, W. Lin, Z. Lu, X. Lin, S. Rahardja, E. Ong, and S. Yao, "Rate control for videophone using local perceptual cues," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 4, pp. 496–507, Apr. 2005.
[7] Y. Liu, Z. G. Li, and Y. C. Soh, "Region-of-interest based resource allocation for conversational video communication of H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 134–139, Jan. 2008.
[8] R.-L. Hsu, M. Abdel-Mottaleb, and A. Jain, "Face detection in color images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 696–706, May 2002.
[9] J. Saragihand, S. S. Lucey, and J. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proc. ICCV*, 2009, pp. 1034–1041.
[10] P. Kortum and W. Geisler, "Implementation of a foveated image coding system for bandwidth reduction of video images," *Proc. SPIE*, vol. 2657, pp. 350–360, 1996.
[11] U. Rauschenbach and H. Schumann, "Demand-driven image transmission with levels of detail and region s of interest," *Comptuer Graphics*, vol. 23, no. 6, pp. 857–866, June 1999.
[12] S. Lee, A. C. Bovik, and Y. Kim, "Low delay foveated visual communications over wireless channels," in *Proc. ICIP*, 1999, pp. 90–94.
[13] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct.. 2004.
[14] T. Liu, N. Zheng, W. Ding, and Z. Yuan, "Video attention: Learning to detect a salient object sequence," in *Proc. ICPR*, 2008.
[15] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image Vis. Comput.*, vol. 29, no. 1, pp. 1–14, Jan. 2011.
[16] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Semantic video analysis for adaptive content delivery and automatic description," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 10, pp. 1200–1209, Oct. 2005.
[17] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[18] O. Hershler and S. Hochstein, "At first sight: A high-level pop out effects for faces," *Vis. Res.*, vol. 45, no. 13, pp. 1707–1724, 2005.
[19] D. Chai and K. N. Ngan, "Face segmentation using skin-color map in videophone applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 4, pp. 551–564, Jun. 1999.
[20] M.-C. Chia, M.-J. Chena, C.-H. Yehb, and J.-A. Jhua, "Region-of-interest video coding based on rate and distortion variations for H.263+," *Signal Process.: Image Commun.*, vol. 23, no. 2, pp. 127–142, Feb. 2008.
[21] G.-L. Wu, Y.-J. Fu, and S.-Y. Chien, "Region-based perceptual quality regulable bit allocation and rate control for video coding applications," in *Proc. VCIP*, 2012.
[22] S. Z. Li and A. K. Jain, *Handbook of Face Recogntition*. New York, NY, USA: Springer, 2011.
[23] G. Boccignone, A. Marcelli, P. Napoletano, G. D. Fiore, G. Iacovoni, and S. Morsa, "Bayesian integration of face and low-level cues for foveated video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 12, pp. 1727–1740, Dec. 2008.
[24] J.-S. Lee, F. D. Simone, and T. Ebrahimi, "Video coding based on audiovisual attention," in *Proc. ICME*, 2009.
[25] M. Nystrom and K. Holmqvist, "Effect of compressed offline foveated video on viewing behavior and subjective quality," *ACM Trans. Multimedia Comput.*, vol. 6, no. 1, pp. 1–14, Jan. 2010.
[26] C.-W. Tang, "Spatiotemporal visual considerations for video coding," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 231–238, Apr. 2007.
[27] J.-S. Lee, F. D. Simone, and T. Ebrahimi, "Efficient video coding based on audio-visual focus of attention," *J. Vis. Commun. Image Represent.*, vol. 24, no. 8, pp. 704–711, Nov. 2011.
[28] JCT-VC, HM 9.0, [Online]. Available: http://hevc.hhi.fraunhofer.de/
[29] H. Choi, J. Yoo, J. Nam, D. Sim, and I. V. Bajic, "Pixel-wise unified rate-quantization model for multi-level rate control," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 6, pp. 1112–1123, Dec. 2013.
[30] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.
[31] 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," *Percept. Motor Skills*, vol. 47, no. 3, pp. 857–858, 1978.
[32] M. T. Pourazad, C. Doutre, M. Azimi, and P. Nasiopoulos, "HEVC: The new gold standard for video compression: how does HEVC compare with H. 264/AVC?," *IEEE Consumer Electron. Mag.*, vol. 1, no. 3, pp. 36–46, Jul. 2012.
[33] Y. Liu, Z. Li, and Y. C. Soh, "A novel rate control scheme for low delay video communication of H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 1, pp. 68–78, Jan. 2007.
[34] "Methodology for the subjective assessment of the quality of television pictures," ITU, Geneva, Switzerland, BT. 500-11, International Telecommunication Union, 2002, pp. 53–56.

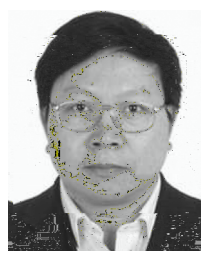


(M'10) received the B.S. degree from Beihang University in 2003, the M.S. degree from Tsinghua University in 2006 and the Ph.D. degree from Imperial College London in 2010. From 2010–2012, he was working as a research fellow at the Electrical Engineering Department, Tsinghua University. Since Jan. 2013, he has been with Beihang University as an Associate Professor. His research interests mainly include visual communication and image processing. He has published more than 30 technical papers in international journals and conference proceedings.



(S'14) received her B.E. degree in electronic engineering from Beihang University, Beijing, China, in June 2013. She is currently a graduate student of Beihang University. During her study in Beihang University, she has won the National Scholarship of Chinese college students and the second prize of national mathematics competition. Her research interests include perceptual video coding and video quality metrics.

(S'14) is now an undergraduate student in School of Electronic and Information Engineering, Beihang University (expected to obtain the B.S. degree in July 2014). He is admitted and will start as a graduate student in Beihang University. During his study, he was awarded the Beihang Gold Medal Honor in 2013, which is the highest honor in Beihang University during the undergraduate period. He was also awarded 15 scholarships, including National Scholarship, and nearly 20 competition prizes. His research interests include rate distortion theory and perceptual video coding.



(M'14) received the B.S. and M.S. degrees in electronic engineering from Beihang University, in 1986 and 1989, respectively. He also received his Ph.D. degree at the same university in 2000. He is currently the dean of school of electronic and information engineering, at Beihang University, Beijing, China. His research interests include image processing, video coding, high-speed signal processing, electromagnetic countermeasure, complex object test, and satellite communications technology. He is author or co-author of over 100 papers and holds 6 patents, as well as published 2 books in these fields. He has undertaken approximately 30 projects related to image/video coding, wireless communication, and etc. Now he has taught "image signal processing" course to undergraduates and "digital signal processing" course to postgraduates for nearly one decade. He is also the expert of China 863 program and the independent director of China Electronic Limited by Share Ltd.