Signal Processing: *Image Communication* 38 (2015) 127–140

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Weight-based R-λ rate control for perceptual HEVC coding on conversational videos ☆

Shengxi Li [a], Mai Xu [a,b,*], Xin Deng [a], Zulin Wang [a]

[a] *School of Electronic and Information Engineering, Beihang University, Beijing 100191, China*
[b] *EDA Lab, Research Institute of Tsinghua University in Shenzhen, Shenzhen, China*

ARTICLE INFO

Available online 14 May 2015

Keywords:
HEVC
Perceptual video coding
Rate control

ABSTRACT

This paper proposes a novel weight-based R-λ scheme for rate control in HEVC, to improve the perceived visual quality of conversational videos. For rate control in HEVC, the conventional R-λ scheme is based on bit per pixel (bpp) to allocate bits. However, bpp does not reflect the visual importance variation of pixels. Therefore, we propose a novel weight-based R-λ scheme to consider this visual importance for rate control in HEVC. We first conducted an eye-tracking experiment on training videos to figure out different importance of background, face, and facial features, thus generating weight maps of encoding videos. Upon the weight maps, our scheme is capable of allocating more bits to the face (especially facial features), using a new term bit per weight. Consequently, the visual quality of face and facial features can be improved such that perceptual video coding is achieved for HEVC, as verified by our experimental results.

1. Introduction

Supported by the recent advances in related techniques, the popularity of multimedia applications has been considerably increased. It has been pointed out in [1] that video applications with high resolutions, such as FaceTime and Skype, occupy a large proportion of data among the existing multimedia applications. The limited bandwidth issue thus becomes more and more serious, causing "spectrum crush". To better relieve the bandwidth-hungry issue, high efficiency video coding (HEVC) standard [1], also called H.265, has been formally established.

Rate control is a crucial module in HEVC, whose aim is to optimize visual quality via reasonably allocating bits to various frames and blocks, at a given bit-rate. An excellent rate control scheme is able to precisely allocate bits, and to output better visual quality of compressed videos. In other words, at the same visual quality, a better rate control scheme consumes less bit-rate and therefore achieves the goal of relieving the bandwidth bottleneck. There are many rate control schemes for different video coding standards (e.g. TM5 for MPEG-2 [2], VM8 for MPEG-4 [3] and JVT-N046 [4] for H.264). For HEVC, a pixel-wise unified rate quantization (URQ) scheme has been proposed in [5] to compute quantization parameter (QP) at a given target bit-rate. Since this scheme works at pixel level, it can be easily applied to blocks with various sizes. However, according to [6], Lagrange multiplier λ [7], which represents the bit cost of encoding a block, is more important than QP in allocating bits. Therefore, a new scheme, R-λ scheme, was proposed in [6] to better allocate the bits in HEVC.

Nevertheless, high resolution video delivery, especially at low bit-rate scenarios, still poses a great challenge to HEVC. In fact, according to the human visual system (HVS), there exists much perceptual redundancy that can be further exploited to greatly improve the coding efficiency of HEVC,

☆ This work was supported by NSFC under Grant no. 61202139 and 61471022 and China 973 program under Grant no. 2013CB29006.

* Corresponding author.
E-mail address: MaiXu@buaa.edu.cn (M. Xu).

http://dx.doi.org/10.1016/j.image.2015.04.011
0923-5965/

thus relieving the bandwidth-hungry issue [8]. For instance, when a person looks at a video, a small region around a point of fixation, called region-of-interest (ROI), is concerned most [8] with high resolution, while the peripheral region is captured at a low resolution. Hence, in light of this phenomenon, a large amount of bits can be saved via reducing perceptual redundancy in the peripheral region, with little loss of perceived quality. Consequently, along with the development of the understanding of the HVS, perceptual video coding is able to more efficiently condense video data.

Rate control for perceptual video coding has received a great deal of research effort from 2000 onwards, due to its great potential in improving coding efficiency [9–12]. In H.263, a perceptual rate control (PRC) scheme [9] was proposed. In this scheme, a perceptual sensitive weight map of conversational scene (i.e., scene with frontal human faces) is obtained by combining stimulus-driven (i.e., luminance adaptation and texture masking) and cognition-driven (i.e., skin colors) factors together. According to such a map, more bits are allocated to ROIs by reducing QP values in these regions. Afterwards, for H.264/AVC, a novel resource 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. Moreover, Xu et al. [13] utilized a novel window model to characterize the relationship between the size of window and variations of picture quality and buffer occupancy, ensuring a better perceptual quality with less quality fluctuation. This model was advanced in [14] with an improved video quality metric for better correlation to the HVS. Most recently, in HEVC the perceptual model of structural similarity (SSIM) has been incorporated for perceptual video coding [15]. Instead of minimizing mean squared error (MSE) and sum of absolute difference (SAD), SSIM is minimized [15] to improve the subjective quality of perceptual video coding in HEVC. However, as pointed out by [16], assigning pixels with weights according to visual attention is much more accurate than SSIM for evaluating the subjective quality. To this end, a scheme [12] was proposed to improve the visual quality and meanwhile to reduce the encoding complexity, via considering the visual attention on ROIs (e.g., face and facial features). However, to our best knowledge, although larger weights are imposed on ROIs in the above approaches, their values are assigned in an arbitrary manner. Moreover, there is no perceptual approach for the latest R-λ rate control scheme [6] in HEVC.

Therefore, we propose a novel weight-based R-λ rate control scheme to improve the perceived visual quality of compressed conversational videos, based on the weights of face regions and facial features learned from eye-tracking data. To be more specific, similar to [12], we consider face regions as ROIs, and also consider facial features (e.g., mouth and eyes) as the most important ROIs. Different from [12], the weights allocated to background, face, and facial features are more precise and reasonable, as they are obtained upon the saliency distribution learnt from our eye-tracking data of several training videos. Based on these weights, the weight-based R-λ rate control scheme is proposed, using a new term, bit per weight (bpw), to enhance the quality of face regions, especially the facial features. Since the perceptual video coding is the main goal of our scheme, we review it in the following.

2. The related work on perceptual video coding

Generally speaking, main parts of perceptual video coding are perceptual models, perceptual model incorporation in video coding and performance evaluations, as illustrated in Fig. 1. Specifically, perceptual models, which imitate the output of the HVS to specify the ROIs and non-ROIs, need to be designed first for perceptual video coding. Secondly, on the basis of the perceptual models and existing video coding standards, perceptual model incorporation in video coding from perceptual aspects needs to be developed to encode/decode the videos, mainly by moving their perceptual redundancy. Rather than incorporating perceptual model in video coding, some machine learning based image/video

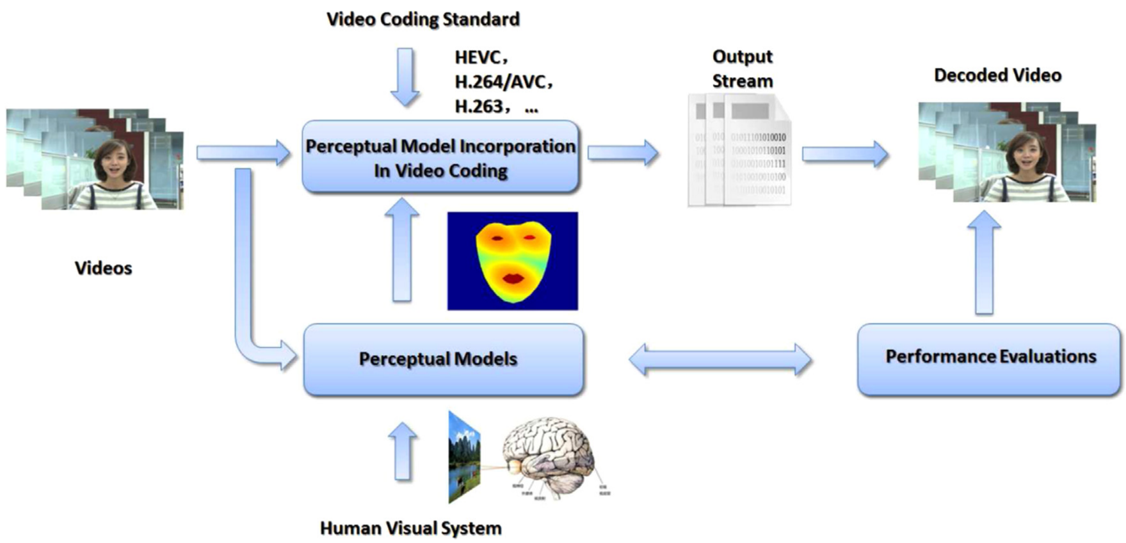

Fig. 1. The framework of perceptual video coding.

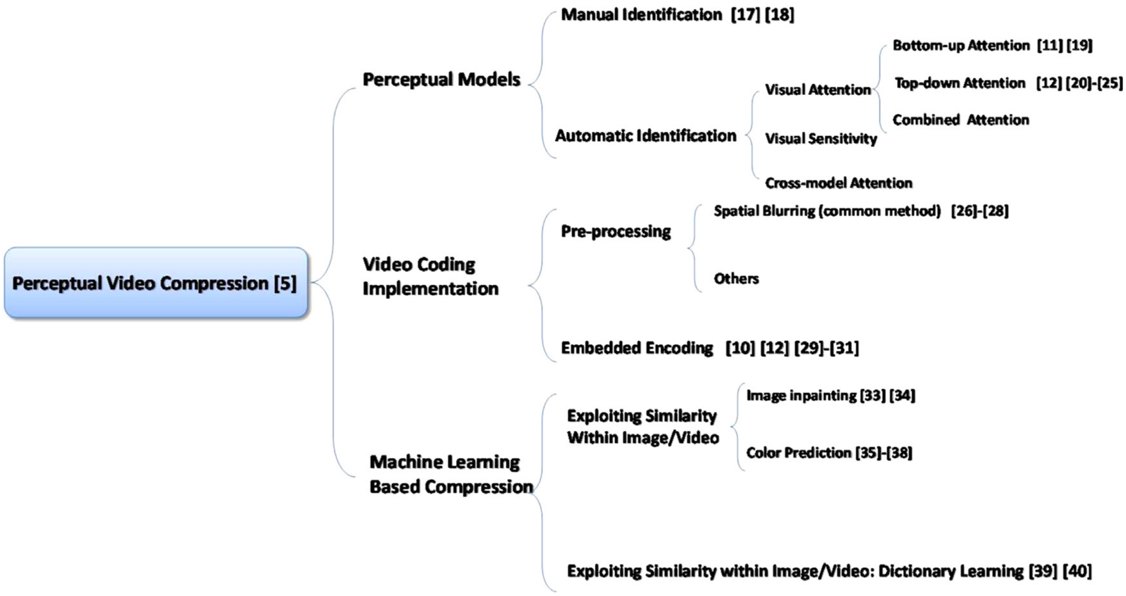

Fig. 2. The literatures on perceptual video coding.

compression approaches have also proposed during the past decade. A summarized literature review is depicted in Fig. 2, which is to be explained in detail in the next two subsections.

2.1. Perceptual model

Perceptual models can be classified into two categories: manual and automatic identification.

2.1.1. Manual identification

This kind of perceptual models requires manual effort to distinguish important regions which need to be encoded with high quality. In the early years, Geisler and Perry [17] employed a foveated multi-resolution pyramid (FMP) video encoder/decoder to compress each image of varying resolutions into 5 or 6 regions in real-time, using a pointing device. This model requires the users to specify which regions attract them most during the video transmission. Thus, this kind of models may lead to transmission and processing delay between the receiver and transmitter sides, when specifying the ROIs. Another way [18] is to specify ROIs before watching, hence avoiding the transmission and processing delay. However, considering the workload of humans, these models cannot be widely applied to various videos.

In summary, the advantage of manual identification models is the accurate detection of ROIs. However, as the cost, it is expensive and intractable to extensively apply these models due to the involvement of manual effort or hardware support. In addition, for the models of user input-based selection, there exists transmission and processing delay, thus making the real-time applications impractical.

2.1.2. Automatic identification

Just as its name implies, this category of perceptual models is to automatically recognize ROIs in videos, according to visual attention mechanisms. Therefore, visual attention models are widely used among various perceptual models. There are two classes of visual attention models: either bottom-up or top-down models. Itti's model [19] is one of the most popular bottom-up visual attention models in perceptual video coding. Mimicking processing in primate occipital and posterior parietal cortex, Itti's model integrates low-level visual cues, in terms of color, intensity, orientation, flicker, and motion, to generate a saliency map for selecting ROIs [11].

The other class of visual attention models is top-down processing [20–25,12]. The top-down visual attention models are more frequently applied to video applications, since they are more correlated with human attractiveness. For instance, human face [10,12,21] is one of the most important factor that draw top-down attention, especially for conversational video applications. Also, a hierarchical perceptual model of face [12] has been established, endowing unequal importance within the face region. However, the above-mentioned approaches are unable to figure out the importance of face region.

In this paper, we quantify the saliency of face and facial features via learning the saliency distribution from the eye fixation data of training videos, via conducting the eye-tracking experiment. Then, after detecting face and facial features for automatically identifying ROI [12], the saliency map of each frame of encoded conversational video is assigned using the learnt saliency distribution. Although the same ROI is utilized in [12], the weight map of our scheme is more reasonable for the perceptual model for video coding, as it is in light of learnt distribution of saliency over face regions. Note that the difference between ROI and saliency is that the former refers to the place that may attract visual attention while the later refers to the possibility of each pixel/region to attract visual attention.

2.2. Perceptual model incorporation in video coding

After setting up the perceptual model, the next task is to apply it in the existing video coding approaches. One category of approaches called pre-processing is to control the non-uniform distribution of distortion before encoding [26–28]. A common way of pre-processing is spatial blurring [26,27]. For instance, the spatial blurring approach [26] separates the scene into foreground and background. The background is blurred to remove high frequency information in spatial domain so that less bits are allocated to this region. However, this may cause obvious boundaries between the background and foreground.

Another category is to control the non-uniform distribution of distortion during encoding, therefore called embedded encoding [29,10,30,31,12]. As it is embedded into the whole coding process, this category of approaches is efficient in more flexibly compressing videos with different demands. In [10], Liu et al. established importance map at the macro block (MB) level based on face detection results. Moreover, combining texture and nontexture information, a linear rate–quantization (R–Q) model is applied to H.264/AVC. Based on the importance map and R–Q model, the optimized QP values are assigned to all MBs, which enhances the perceived visual quality of compressed videos. In addition, after obtaining the importance map, the other encoding parameters, such as mode decision and motion estimation (ME) search, are adjusted to provide ROIs with more encoding resources. Xu et al. [12] proposed a new weight-based URQ rate control scheme for compressing conversational videos, which assigns bits according to bpw, instead of bit per pixel (bpp) in conventional URQ scheme. Then, the quality of face regions is improved such that its perceived visual quality is enhanced. The scheme in [12] is based on the URQ model [5], which aims at establishing the relationship between bite-rate R and quantization parameters Q, i.e., R–Q relationship. However, since various flexible coding parameters and structures are applied in HEVC, R–Q relationship is hard to be precisely estimated [32]. Therefore, Lagrange multiplier λ [7], which stands for the slope of R–D curve, has been investigated. According to [32], the relationship between λ and R can be better characterized in comparison with R–Q relationships. This way, on the basis of R-λ model, the state-of-the-art R-λ rate control scheme [6] has better performance than the URQ scheme. Therefore, on the basis of the latest R-λ scheme, this paper proposes a novel weight-based R-λ scheme to further improve the perceived video quality of HEVC.

2.3. Machine learning based compression

From the viewpoint of machine learning, the pixels or blocks from one image or several images may have high similarity. Such similarity can be discovered by machine learning techniques, and then utilized to decrease redundancy of video coding. For exploiting the similarity within an image/video, image inpainting has been applied in [33,34] to use the image blocks from spatial or temporal neighbors for synthesizing the unimportant content, which is deliberately deleted at the encoder side. As such, the bits can be saved as not encoding the missing areas of the image/video. Also, rather than predicting the missing intensity information in [33,34] , several approaches [35–38] have been proposed to learn to predict the color in images using the color information of some representative pixels. Then, only representative pixels and gray scale image need to be stored, such that the image [38,36,37] or video [35] coding can be achieved.

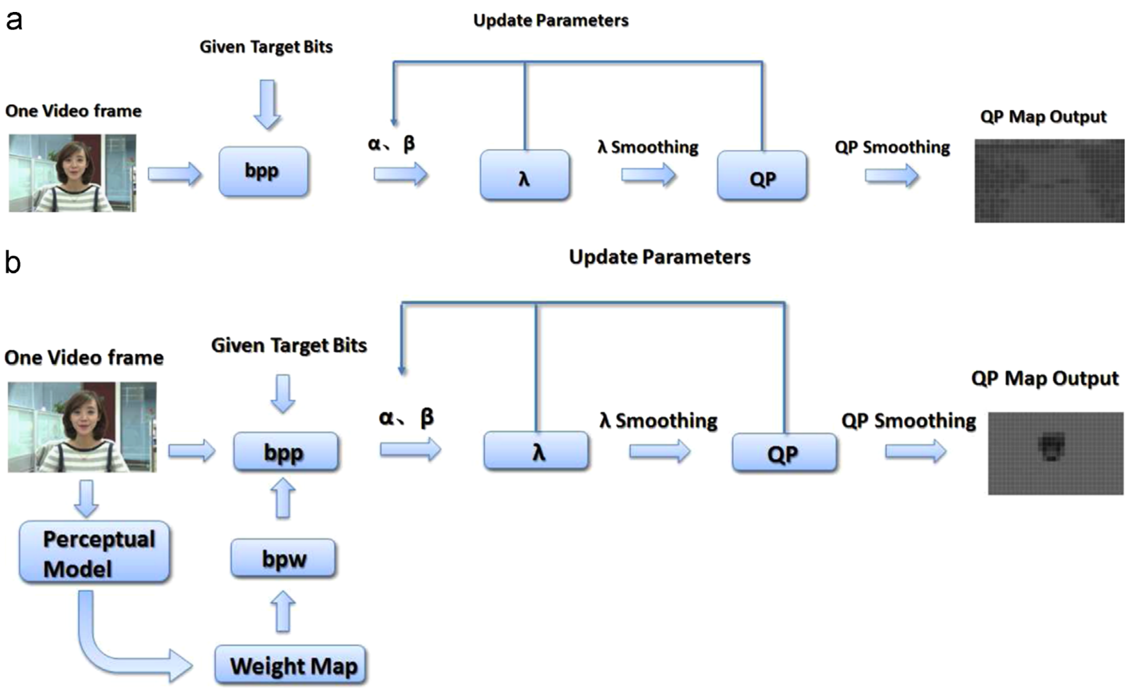

Fig. 3. (a) The procedure of the conventional R-λ and (b) our rate control schemes.

For working on similarity across various images or frames of videos, dictionary learning has been developed to discover the inherent patterns of image blocks. Together with dictionary learning, sparse representation can then be used to effectively represent an image for image [39] or video coding [40], instead of conventional image transforms such as discrete cosine transform (DCT).

3. Review of HEVC R-λ scheme

The main goal of rate control in video coding is minimizing the distortion of a compressed video at a given bit-rate. In order to better achieve this goal, as illustrated in Fig. 3(a), R-λ rate control scheme [6] calculates a Lagrange multiplier λ before computing QP. From Fig. 3(a), we can also see that the main steps of this scheme are working out the bpp-λ and λ-QP relationships, to finally output QP values. So, we review bpp-λ and λ-QP relationships in the following for the R-λ rate control scheme. Note that this paper only focuses on the rate control at largest coding unit (LCU) level.

3.1. bpp-λ relationship

Parameter λ, which is the slope of rate–distortion (R–D) curve [6] (also seen as Lagrange multiplier), is crucial during the rate control process. The relationship between λ and R–D curve can be formulated by

$$\lambda = -\frac{\partial D}{\partial R}, \tag{1}$$

where D and R represent the distortion and bit-rate for one LCU.

Furthermore, the Hyperbolic model $D = CR^{-K}$, characterizing the relationship between R and D, is adopted in the rate control scheme [7,41]. Here, C and K are parameters determined by the characteristic of video content. Then, with (1), the relationship between R-λ can be obtained by

$$\lambda = -\frac{\partial D}{\partial R} = CK \cdot R^{-K-1} = a \cdot R^{b}, \tag{2}$$

where $a = CK$ and $b = -K-1$ are parameters related to the video content as well. As different LCUs are of different contents, a and b need to be updated along with the encoding process of each LCU.

Next, once R is obtained for an LCU, λ can be an output for estimating QP of the current processing LCU. Here, the bit-rate R can be modeled in terms of bpp by

$$R = \text{bpp} \cdot f \cdot w \cdot h, \tag{3}$$

where w and h are the width and height of the video frame; f represents the frame rate. Recall that bpp is the bit per pixel for this LCU. Upon (3), (2) can be rewritten as

$$\lambda = \alpha \cdot \text{bpp}^{\beta}, \tag{4}$$

where $\alpha = a \cdot (fwh)^{b}$ and $\beta = b$ are parameters also related to video contents. In HEVC, α and β need to be updated during encoding, with their initial values set to 3.2003 and -1.367, i.e., after encoding LCU, α and β are updated for the co-located LCU of the subsequent frames. Note that as shown in [32], different initial values of α and β have little impact on the compressed videos, both on R–D performance and bit-rate error. Assuming that the actual encoded bpp is $\check{\text{bpp}}$ and the actual used λ is $\check{\lambda}$ for the current encoding, α and β can be updated by α' and β':

$$\alpha' = \alpha + \delta_{\alpha} \cdot \left(\ln \check{\lambda} - \ln(\alpha \cdot \check{\text{bpp}}^{\beta}) \right) \cdot \alpha,$$
$$\beta' = \beta + \delta_{\beta} \cdot \left(\ln \check{\lambda} - \ln(\alpha \cdot \check{\text{bpp}}^{\beta}) \right) \cdot \ln \check{\text{bpp}}, \tag{5}$$

where δ_{α} and δ_{β} are constants empirically set to 0.1 and 0.05 [6], respectively. Note that $\check{\text{bpp}}$ means the actually consumed bpp after encoding each LCU, while $\check{\lambda}$ represents the actual λ used for calculating QP during encoding each LCU. In general, $\check{\lambda}$ is not equal to $\alpha \cdot \check{\text{bpp}}^{\beta}$ since α and β are unable to accurately fit the relationship between distortion and bit-rate for each LCU. The proof of the updating way can refer to the appendix of [32].

Then, once bpp value is achieved for each LCU, λ can be estimated with (4). Here, we assume that bpp and λ values for the j-th LCU are bpp_j and λ_j, respectively. Next, assuming that the number of pixels in the j-th LCU is N_j, we obtain bpp_j through the target bits T_j for the j-th LCU:

$$\text{bpp}_j = \frac{T_j}{N_j}, \tag{6}$$

and

$$T_j = \frac{\hat{T} - B}{\sum_{i=j}^{M-1} c_i} \cdot c_j, \tag{7}$$

where $\hat{T}$ is the target bits remaining for encoding this frame and B is the remaining header bits. Moreover, there are M LCUs in this frame, and c_i means the texture complexity of the i-th LCU. To be more specific, for inter frames, the target bits are allocated according to MAD of the co-located LCU in the previous pictures, i.e., c_i is related to MAD for the i-th LCU. For intra frames, c_i is related to the sum of absolute transformed difference (SATD) of the i-th LCU [42]. For computing c_i, see [43].

In addition, before establishing λ-QP relationship, there exists a step to smooth λ with value $\tilde{\lambda}_j$ for the j-th LCU:

$$\tilde{\lambda}_j = \max\{\max\left(\lambda_P \cdot 2^{\frac{-2.0}{3.0}}, \tilde{\lambda}_{j-1} \cdot 2^{\frac{-1.0}{3.0}}\right),$$
$$\min\{\lambda_j, \min\left(\lambda_P \cdot 2^{\frac{2.0}{3.0}}, \tilde{\lambda}_{j-1} \cdot 2^{\frac{1.0}{3.0}}\right)\}\}, \tag{8}$$

where λ_P represents the λ value of the current frame. $\tilde{\lambda}_{j-1}$ represents λ value which has been smoothed in the $(j-1)$-th LCU . The calculation of λ_P is depicted in [6].

3.2. λ-QP relationship

After establishing bpp-λ relationship, the remaining task is finding out λ-QP relationship. Mathematically, QP value can be obtained through a multiple-QP optimization process:

$$\min J(\text{QP}) = D(\text{QP}) + \lambda R(\text{QP}), \tag{9}$$

to provide the smallest RD cost $J(\text{QP})$, with distortion $D(\text{QP})$ and rate $R(\text{QP})$.

The optimal QP can be achieved as the final output of rate control via solving (9). However, this optimization hugely increases encoding complexity. To reduce the encoding complexity, a fitting formulation, rather than multiple-QP optimization, is proposed [44,45] to determine QP value QP_j for the j-th LCU:

$$QP_j = \theta_0 \cdot \ln \tilde{\lambda}_j + \theta_1. \tag{10}$$

Recall that $\tilde{\lambda}_j$ is the smoothed λ value of the j-th LCU. In (10), θ_0 and θ_1 are coefficients fitting the linear relationship between QP and $\ln \lambda$. Note that θ_0 and θ_1 are empirically set to 4.2005 and 13.7122, respectively, in [44]. Their values remain the same during each video coding. Similar to (8), QP_j needs to be smoothed as well:

$$\tilde{QP}_j = \max\{\max(\tilde{QP}_{j-1} - 1, QP_P - 2), \\ \min\{QP_j, \min(\tilde{QP}_{j-1} + 1, QP_P + 2)\}\}, \tag{11}$$

where QP_P is the QP value of the current frame. For the calculation of QP_P, please refer to [6]. Moreover, $\tilde{QP}_{j-1}$ means the smoothed QP value of the $(j-1)$-th LCU. Finally, the $\tilde{QP}_j$ can be output in R-λ rate control scheme.

From the definition of bpp, it is worth pointing out that each pixel is endowed with the equal visual importance over the whole video frame. Therefore, the R-λ scheme wastes many bits on encoding the non-ROIs, to which humans pay less attention.

4. The proposed rate control scheme

This section proposes the weight-based R-λ rate control scheme, to take into account local visual importance of video content. Fig. 3(b) shows the procedure of our weight-based R-λ rate control scheme. Specifically, we first establish a perceptual model of face by learning from eye fixation points of our eye-tracking experiment. Note that such a perceptual model is learnt offline from training videos that are different from encoding videos. Based on such a perceptual model, the weight-based R-λ rate control scheme is proposed to improve the visual quality of face and facial features, thus providing a better perceived quality.

4.1. Learning for perceptual model

In [8], the authors have shown that face draws a majority of attention in conversational videos. It is interesting to further quantify unequal importance of background, face and facial features to human attention. In this section, we conducted eye tracking experiment on the training conversational videos to obtain values of such unequal importance so that these values can be used for encoding other videos.

Before the experiment, it is necessary to first extract the face and its facial features in conversational videos using the method of [12]. Generally speaking, our extraction technique is based on a real-time face alignment method [46]. To be more specific, several key landmarks obeying the point distribution model (PDM) are located in the face of an image using the method in [46], which combines the local detection (texture information) and global optimization (facial structure) together. Here, 66 landmarks, produced by the PDM, are connected to precisely identify the contours and regions of face and facial features. Note that the extraction in our scheme can be achieved in real-time, as the face alignment method [46] is indeed fast. Also, the 3000 fps face alignment [47] may be used to further speed up the extraction on face and its facial features.

For the eye tracking experiment, 18 conversational video clips (resolution: 720×480) were collected and each of them was cut to 750 frames at 25 Hz. Note that these conversational video clips were collected from movies, news, and videos captured by Nikon D800 camera. Also, note that all training videos are different from the test videos of Section 5. These video clips were then presented at a random order to 24 subjects (14 males, 10 females, aging 22–32). The subjects were seated on an adjustable chair at a viewing distance of 60 cm, ensuring that the subject's horizontal sight is in the center of screen. The eye fixation points of all subjects were recorded over frames of each clip by Tobii T60 eye tracker. Some of the recorded eye tracking data are available at our website http://www.ee.buaa.edu.cn/xumfiles. One example of eye tracking results is shown in Fig. 4. Next, we focus on quantifying the visual attention on different regions of conversational videos by combining the eye fixation points of all subjects together.

After the eye tracking experiment, f_r, f_l, f_m, f_n, f_o, and f_b, which denote eye fixation points of all subjects falling into right eye, left eye, mouth, nose, other parts in face and background, were counted. Given the counted eye fixation points (efp) of different regions, we have the degrees of

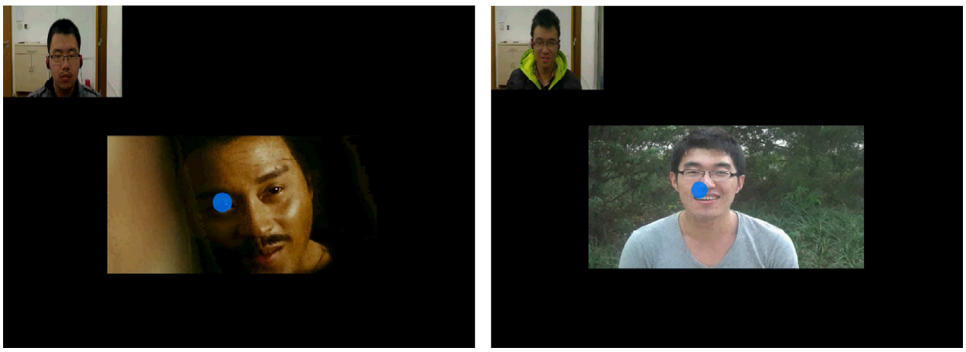

Fig. 4. Example of eye tracking results. The blue circles show the positions of eye fixation points and their sizes represent the staying duration of eye fixation points.

Table 1
The values of visual attention degrees of different regions.

	c_r	c_l	c_m	c_n	c_o	c_b
efp/p	0.122	0.108	0.116	0.080	0.076	0.002

visual attention in these regions:

$$\begin{cases} c_r = f_r/p_r \\ c_l = f_l/p_l \\ c_m = f_m/p_m \\ c_n = f_n/p_n \\ c_o = f_o/p_o \\ c_b = f_b/p_b, \end{cases} \tag{12}$$

where p_r, p_l, p_m, p_n, p_o, and p_b are defined as the numbers of pixels in the regions of right eye, left eye, mouth, nose, other parts in face and background. c_r, c_l, c_m, c_n, c_o, and c_b are the values of visual attention degrees for these regions. Their values output by our eye tracking experiment are reported in Table 1. Note that the degrees of visual attention for face and facial features may vary in different videos, according to video content. However, the values in Table 1 are constant to simply predict the visual attention attended to face and facial features, as it is hard to take into account the attention variation of face and facial features, caused by different video content.

Finally, assuming that the background weight is 1, the weight map of a video frame can be computed upon the results of Table 1 by

$$w_n = \begin{cases} 1 & \text{if } n \in \text{background} \\ c_r/c_b & \text{if } n \in \text{right eye} \\ c_l/c_b & \text{if } n \in \text{left eye} \\ c_m/c_b & \text{if } n \in \text{mouth} \\ c_n/c_b & \text{if } n \in \text{nose} \\ c_o/c_b & \text{if } n \in \text{others}, \end{cases} \tag{13}$$

where w_n is the weight of the n-th pixel in the video frame. Given (13), we can obtain the weight map for each conversational video frame to be encoded, with extracted face and facial features [12].

4.2. Weight-based R-λ rate control scheme

For the conventional R-λ rate control scheme, bpp is an important term to allocate bits. In our scheme, rather than replacing bpp in [12], a new term bpw is introduced to calculate bpp, for allocating bits in accordance with the weight map of our perceptual model. Before encoding a frame, we first initialize bpw in order to separate the target bits of the whole frame into the target bits of ROIs (face and facial features) and non-ROIs (background):

$$\overline{\text{bpw}} = \frac{T}{\sum_{n=1}^{N} w_n}, \tag{14}$$

where T stands for the target bits of a frame. Recall that w_n means the weight of the n-th pixel. N is the number of pixels in this frame. Then, the target bits T' for ROIs and T'' for non-ROIs are

$$\begin{cases} T' + T'' = T \\ T'' = \dfrac{\sum_{n \in \mathbf{n}''} w_n \overline{\text{bpw}}}{\sum_{n \in \mathbf{n}'} w_n \overline{\text{bpw}}} T', \end{cases} \tag{15}$$

where $\mathbf{n}'$ denotes the indices of ROI pixels, the weights of which are larger than 1; $\mathbf{n}''$ means the indices of non-ROI pixels, the weights of which are equivalent to 1.

By solving (15), T' and T'' are obtained:

$$T' = \frac{\sum_{n \in \mathbf{n}'} w_n}{\sum_{n \in \mathbf{n}'} w_n + \sum_{n \in \mathbf{n}''} w_n} T, \tag{16}$$

$$T'' = \frac{\sum_{n \in \mathbf{n}''} w_n}{\sum_{n \in \mathbf{n}'} w_n + \sum_{n \in \mathbf{n}''} w_n} T. \tag{17}$$

Then, the target bits for ROIs and non-ROIs can be reasonably arranged, according to the importance of these regions. Next, bpw_j for the j-th LCU can be calculated as

$$\text{bpw}_j = \left\{ \frac{\hat{T}'}{\sum_{n \in \hat{n}'} w_n} \; j \in \mathbf{m}' \frac{\hat{T}''}{\sum_{n \in \hat{n}''} w_n} \; j \in \mathbf{m}'', \right. \tag{18}$$

where $\mathbf{m}'$ and $\mathbf{m}''$ are the LCU indices of ROIs and non-ROIs, respectively. Note that LCU of ROIs means that the average weight of this LCU is larger than 1. Otherwise, LCU belongs to non-ROIs. Besides, $\hat{T}'$ and $\hat{T}''$ denote the remaining target bits for ROIs and non-ROIs; $\hat{n}'$ and $\hat{n}''$ represent the pixel indices of the current and its subsequent LCUs for ROIs and non-ROIs.

Afterwards, T_j, the target bits of the j-th LCU, can be estimated via

$$T_j = \sum_{n \in \mathbf{n}_j} w_n \cdot \text{bpw}_j, \tag{19}$$

where $\mathbf{n}_j$ denotes the pixel indices in the j-th LCU. It can be seen from (19) that the LCU with large bpw and w_n is able to be assigned with more target bits. So, ROIs, especially the more important ROIs (e.g., facial features), are emphasized with more target bits.

Finally, given the newly estimated T_j of (19), bpp_j can be acquired with (6). Then, the rate control can be achieved through (4) and (10) with bpp_j known for each LCU. In addition, we adjust the boundary setting of λ_j and QP_j based on the weights from our perceptual model to impose the ROIs more priority on bit allocation. Specifically, (8) is rewritten as follows:

$$\tilde{\lambda}_j = \max\left\{ \max\left(\tilde{\lambda}_{j-1} \cdot 2^{\frac{-1.0}{3.0}} / \frac{\sum_{n \in \mathbf{n}_j} w_n}{N_j}, \lambda_P \cdot 2^{\frac{-2.0}{3.0}} / \frac{\sum_{n \in \mathbf{n}_j} w_n}{N_j} \right), \right.$$
$$\left. \min\left\{ \lambda_j, \min\left(\tilde{\lambda}_{j-1} \cdot 2^{\frac{1.0}{3.0}} \cdot \frac{\sum_{n \in \mathbf{n}_j} w_n}{N_j}, \lambda_P \cdot 2^{\frac{2.0}{3.0}} \cdot \frac{\sum_{n \in \mathbf{n}_j} w_n}{N_j} \right) \right\} \right\}, \tag{20}$$

and QP boundary smoothing (11) is modified correspondingly:

$$\tilde{\text{QP}}_j = \max\left\{ \max\left(\tilde{\text{QP}}_{j-1} - \frac{\sum_{n \in \mathbf{n}_j} w_n}{N_j}, \text{QP}_P - \frac{2}{N_j} \sum_{n \in \mathbf{n}_j} w_n \right), \right.$$

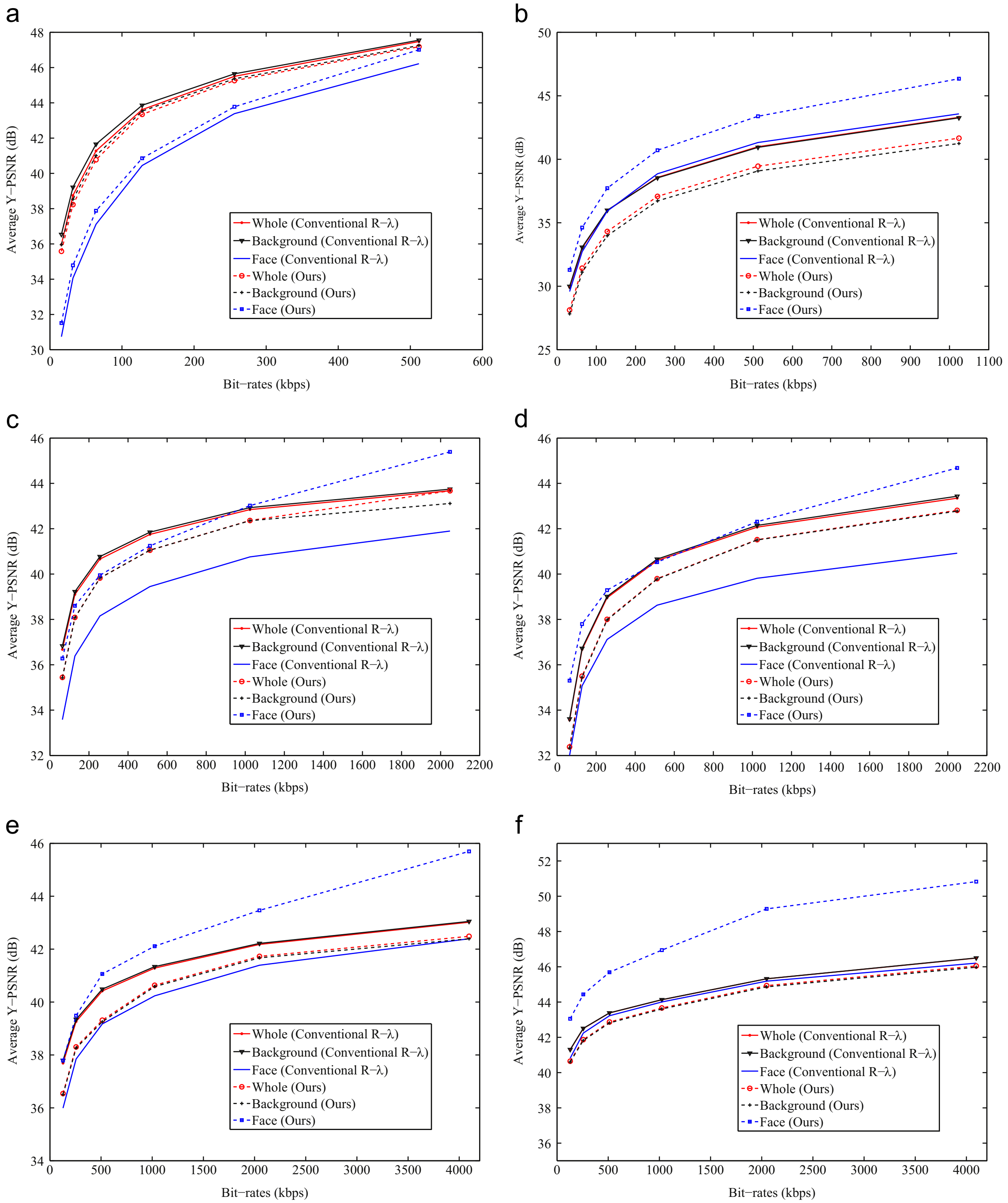

Fig. 5. Rate–distortion performance comparison over face, background, and whole regions between the conventional R-λ and our schemes on compressing six conversational video sequences.

$$\min\left\{\mathrm{QP}_j, \min\left(\tilde{\mathrm{QP}}_{j-1}+\frac{\sum_{n\in \mathbf{n}_j} w_n}{N_j}, \mathrm{QP}_P+\frac{2}{N_j}\sum_{n\in \mathbf{n}_j} w_n\right)\right\}. \tag{21}$$

As can be seen from (20) and (21), more variation is allowed for λ_j and QP_j to improve the quality of ROI regions with more assigned bits. Consequently, the region

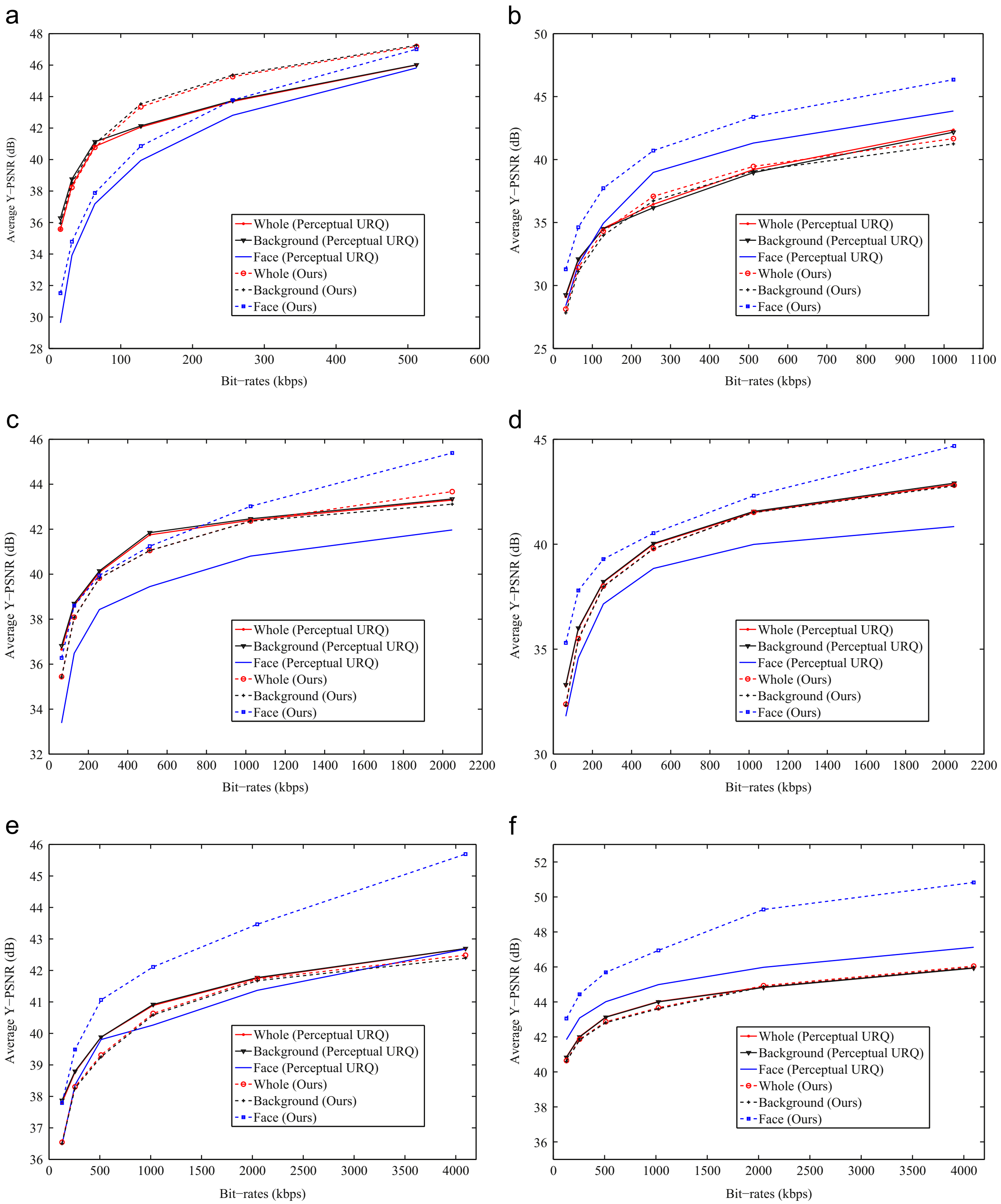

Fig. 6. Rate–distortion performance comparison over face, background, and whole regions between the perceptual URQ [12] and our schemes on compressing six conversational video sequences.

with larger weights has a broader boundary, resulting in better visual quality in ROIs.

In general, we utilize the new term bpw to estimate the target bit for each LCU. Then, bpp is acquired based on bpw, followed by λ and QP values. After encoding one LCU, its QP can be the output. In addition, the relevant parameters, such as α and β, need to be updated for the following LCU. This way, the weight-based R-λ scheme

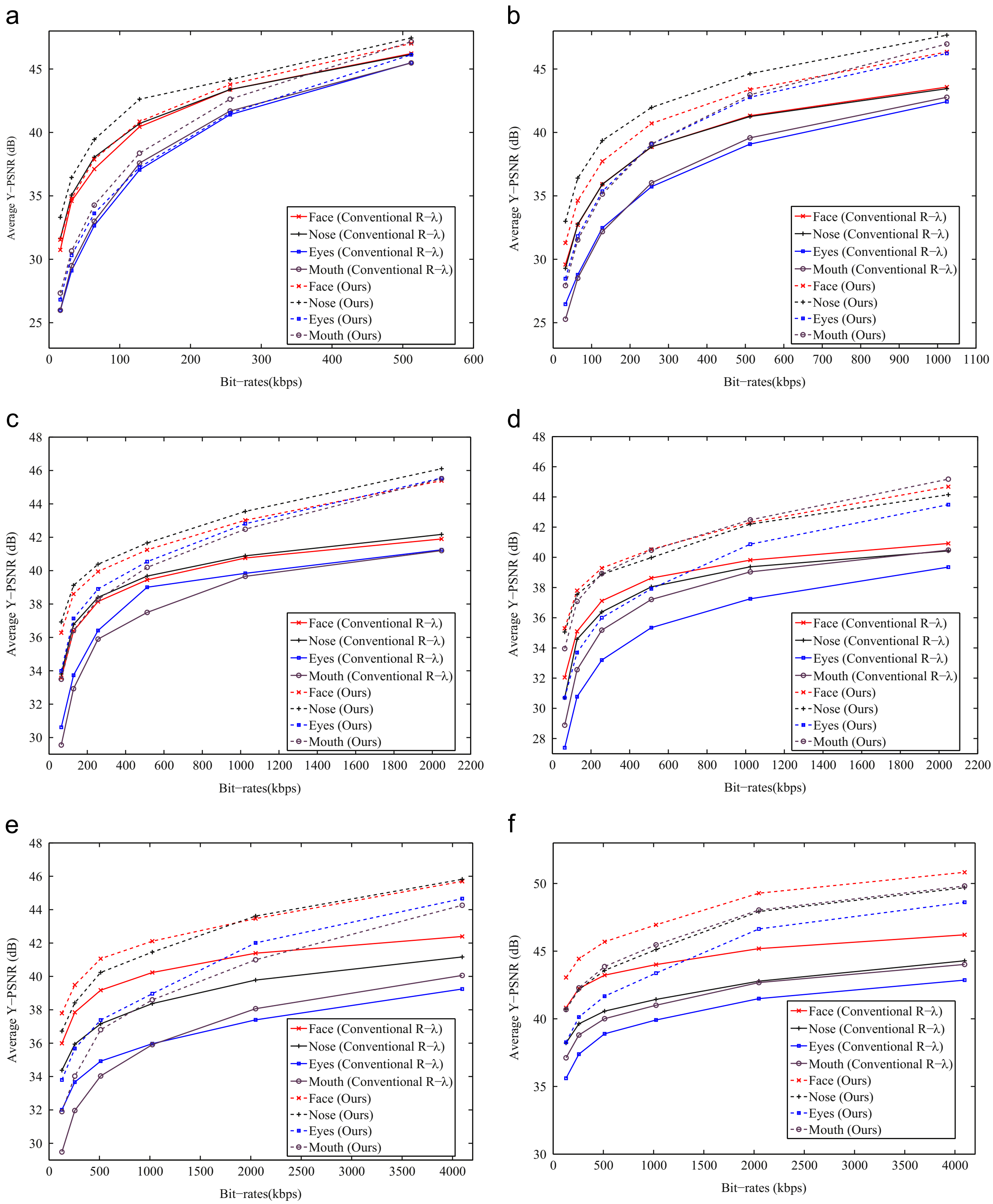

Fig. 7. Rate–distortion performance comparison over facial features between the conventional R-λ and our schemes on compressing six conversational video sequences.

iterates to obtain QP values of each LCU until the last LCU finishes encoding. The main difference between our scheme and the conventional R-λ schemes is that we exploit the weight of each pixel from our perceptual model and bpw to estimate bpp for each LCU. bpp is therefore correspondingly adjusted according to the weights of LCU and bpw. Larger values of weights and bpw, which indicate the more important regions and

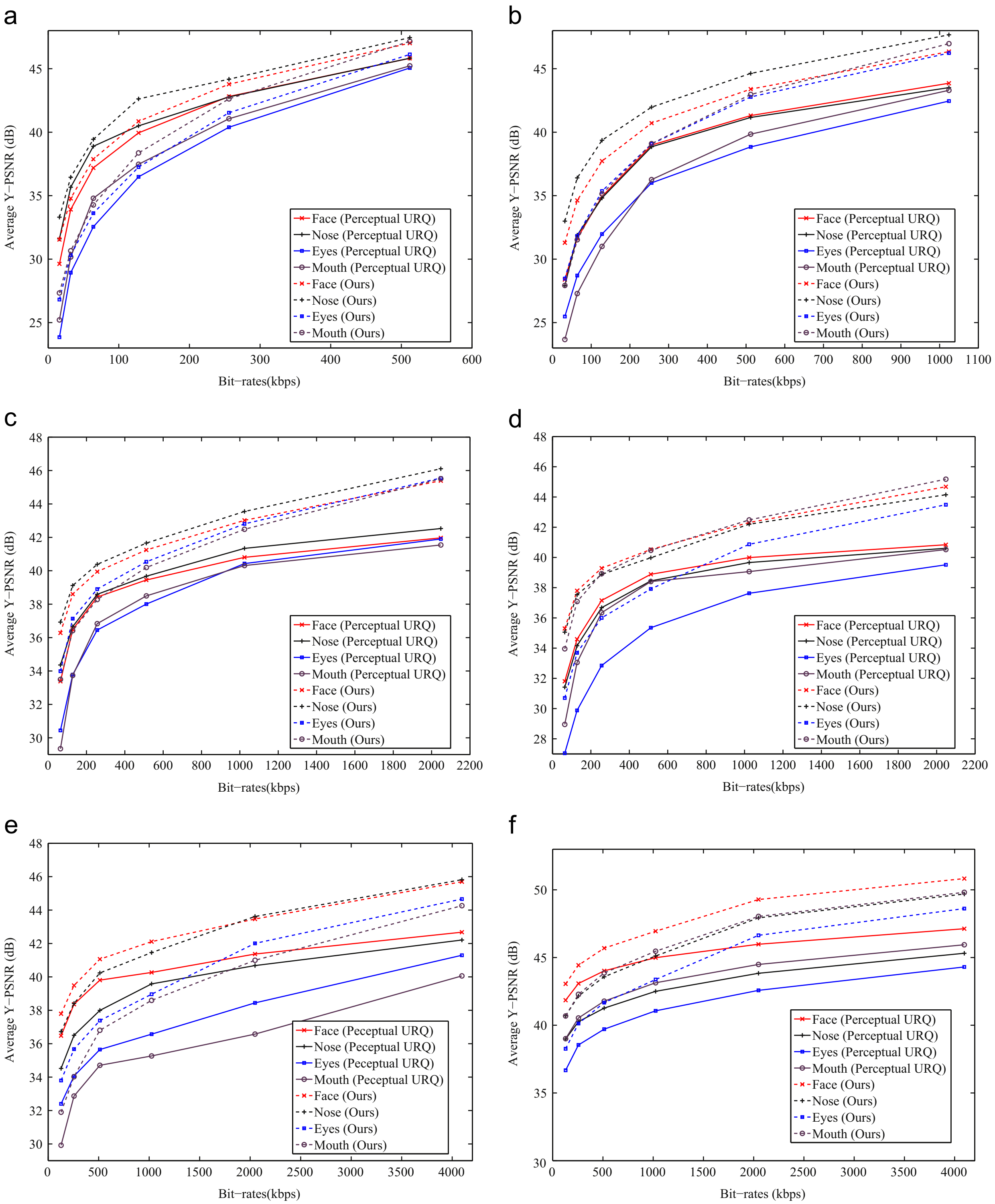

Fig. 8. Rate–distortion performance comparison over facial features between the perceptual URQ scheme [12] and our schemes on compressing six conversational video sequences.

greater bit-rates, lead to larger bpp, thus probably achieving better quality. This way, the ROIs, especially the more important ROIs, are allocated more bits to ensure better perceived quality.

5. Experimental results

In this section, experimental results are presented to validate the proposed weight-based R-λ scheme for

Table 2
DMOS comparison of conventional, perceptual urq, and our schemes.

Sequences	Resolution	Bit-rates (kbps)	Conventional R-λ	Perceptual URQ	Our
Akiyo	352×288	16	55.25	56.34	45.12
		64	29.61	25.59	24.37
Foreman	352×288	32	70.65	72.64	66.58
		128	34.25	36.90	29.86
Johnny	1280×720	64	60.78	59.53	54.54
		256	33.89	29.48	25.69
Vidyo4	1280×720	64	67.16	68.85	61.67
		256	37.21	34.29	27.56
Yan	1920×1080	128	70.64	67.26	65.99
		512	36.35	32.48	29.64
Lee	1920×1080	128	56.40	53.40	47.70
		512	42.06	36.70	29.26

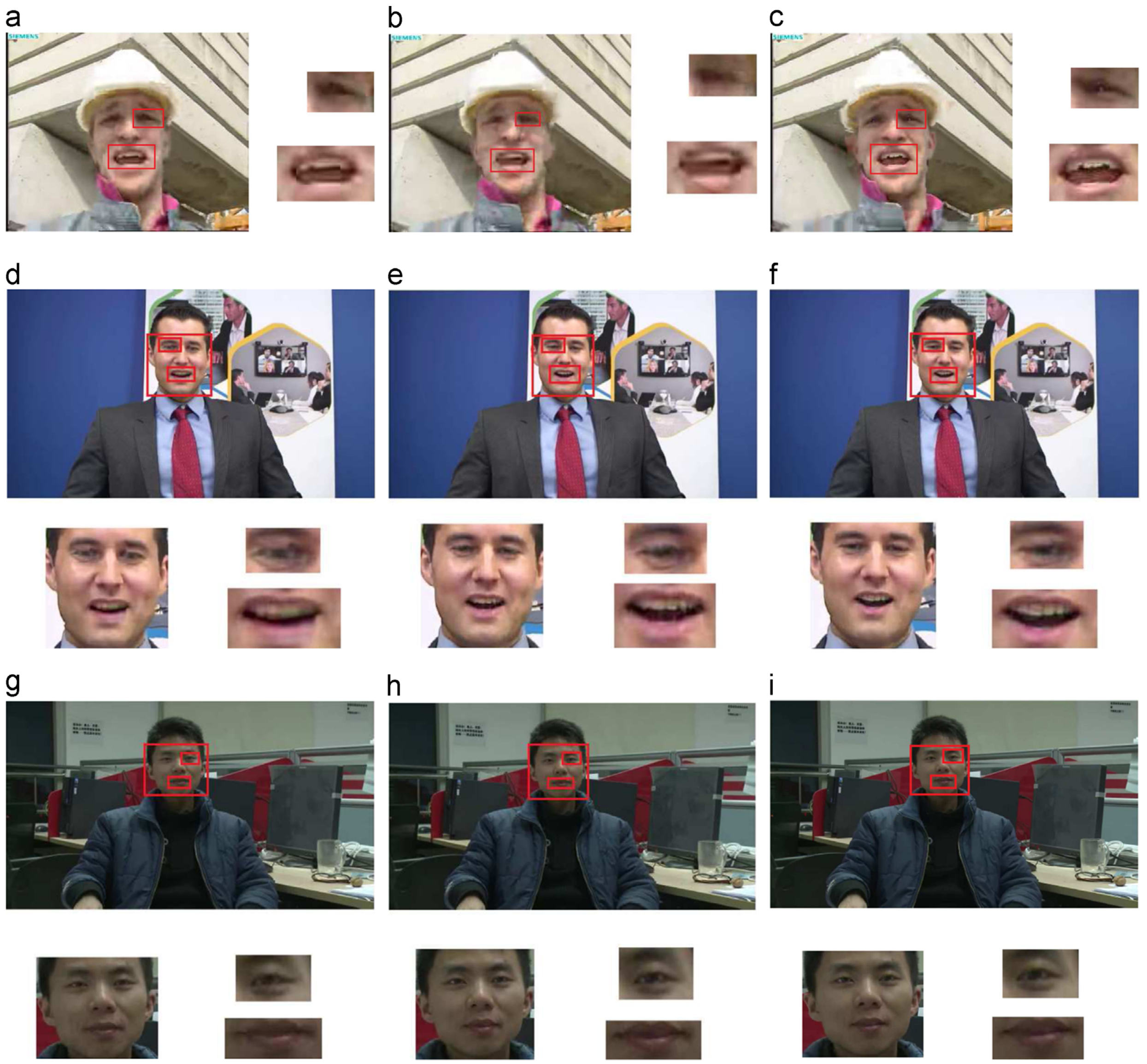

Fig. 9. Visual quality comparison of random selected frames of *Foreman* (CIF resolution), *Johnny* (720p resolution), and *Lee* (1080p resolution). (a), (b), and (c) show the 56th decoded frames of *Foreman* compressed at 32 kbps. (d), (e), and (f) show the 23rd decoded frames of *Johnny* compressed at 64 kbps. Moreover, (g), (h), and (i) show the 41st decoded frames of *Lee* compressed at 128 kbps.

perceptual coding of conversational videos on HEVC platform. We used six test video sequences: two CIF conversational video sequences *Akiyo* and *Foreman*, two 720p conversational videos *Johnny* and *Vidyo4* from HEVC test database, and two 1080p conversational video sequences *Yan* and *Lee* from [12]. We utilized the HEVC test model (HM 16.0 software) with its default R-λ rate control scheme [6] as the reference scheme. Our weight-based R-λ rate control scheme was then embedded into HM 16.0 for comparison. Furthermore, HEVC perceptual video coding work of [12], called perceptual URQ, was used in the comparison. Note that *Lee* was captured in a dark room in order to validate the robustness of our scheme to poor illumination. The parameter setting file *encoder_lowdelay_ P_main.cfg* was used with videos of 150 frames at 25 Hz.

5.1. Objective quality comparison

Figs. 5 and 6 show the rate–distortion performance of the conventional, perceptual URQ, and our schemes in face, background, and whole regions. As can be seen from these figures, our scheme outperforms the conventional R-λ scheme on HM 16.0 platform in terms of average Y-PSNR of the face regions, for all video sequences at different bit-rates. Moreover, the quality of face regions of perceptual URQ scheme is lower than that of our scheme. As the cost, the average Y-PSNR of the background is decreased in our scheme. However, thanks to the HVS, the perceived video quality is increased as verified in the next subsection.

Moreover, Figs. 7 and 8 are plotted to further show the improvement within face regions. One may observe from these figures that the rate–distortion performance of facial features is significantly improved at various bit-rates, for all CIF, 720p, and 1080p videos, in our scheme over both conventional R-λ and perceptual URQ schemes. Furthermore, the quality improvement within a face for the 720p and 1080p videos is much better than that for the CIF videos. This may be due to the fact that more bits can be allocated to ROIs in the 720p and 1080p videos.

5.2. Subjective quality comparison

Since humans are the final receivers of video assessment, subjective evaluation is the most accurate and convincing metric [48]. In this paper, we adopted a single stimulus continuous quality scale (SSCQS) procedure, proposed by Rec. ITU-R BT.500, to rate the subjective quality. The evaluation we conducted was divided into three sessions for CIF, 720p, and HD videos, respectively. Note that the uncompressed reference and test video sequences in each session were displayed with a random order. Before each session, the observers were required to view 5 other training videos (one per quality scale) to help them better understand the subjective quality assessment. 15 observers (5 females and 10 males), aging 19–34, were involved in this test. Note that the observers are different from the subjects in eye-tracking experiment. We used a 24" HP LS24B370 LCD monitor with its resolution being set to 1920×1080 to display the videos. Note that all the videos are displayed in their original resolutions, to avoid the influence of scaling operation. The viewing distance was set to be three to four 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, difference mean opinion scores (DMOS) were computed to reveal the difference of subjective quality between the compressed and uncompressed videos. Smaller value of DMOS corresponds to better subjective quality of the compressed video sequence. Then, Table 2 compares the average DMOS values of all compressed video sequences. From this table, we can see that the DMOS values of our scheme are smaller than the perceptual URQ scheme, and much less than the conventional R-λ scheme, especially at high resolutions. Therefore, our scheme can provide higher subjective video quality. It can be further seen from this table that our scheme is able to perform better than the perceptual URQ scheme at low bit-rates (in comparison with the conventional R-λ scheme). Moreover, the improvement of subjective quality of our scheme over perceptual URQ implies the effectiveness of the learning strategy on allocating weights on face regions. It is because our scheme has better subjective quality, while maintaining comparable Y-PSNRs with perceptual URQ scheme for whole video frames. Fig. 9 further shows the visual quality of our and conventional R-λ schemes.

In summary, our subjective results here, together with the previous objective results, illustrate that our scheme on conversational video coding of HEVC is significantly superior in perceived visual quality.

6. Conclusion

This paper has proposed a novel weight-based R-λ scheme for the rate control of conversational videos in HEVC, to improve its perceived visual quality. First, a perceptual model was established by learning from the training videos with eye fixation points in our eye-tracking experiment, to reveal the importance of visual content for conversational video coding. Then, weight maps can be generated for the encoding video frames. With such maps, a novel weight-based R-λ rate control scheme was proposed for HEVC using bpw to take into account the visual importance of each pixel. Thus, in accordance with HVS, the perceived visual quality is improved by our scheme, as more bits are assigned to ROIs (faces), especially the more important ROIs (facial features). Finally, the experimental results verified such an improvement over several conversational video sequences on HEVC platform (HM 16.0).

References

[1] G.J. Sullivan, J. Ohm, W. Han, T. Wiegand, Overview of the high efficiency video coding (HEVC) standard, IEEE Trans. Circuits Syst. Video Technol. 22 (12) (2012) 1649–1668.

[2] B.G. Haskell, Digital Video: An Introduction to MPEG-2, Springer Science and Business Media, New York, USA, 1997.

[3] A. Vetro, H. Sun, Y. Wang, MPEG-4 rate control for multiple video objects, IEEE Trans. Circuits Syst. Video Technol. 9 (1) (1999) 186–199.

[4] G.J. Sullivan, T. Wiegand, K.-P. Lim, Text description of joint model reference encoding methods and decoding concealment methods,

Document: JVT-N046, Joint Video Team (JVT) of ISO/IECMPEG and ITU-T VCEG.
[5] H. Choi, J. Yoo, J. Nam, D. Sim, I. Bajic, Pixel-wise unified rate-quantization model for multi-level rate control, J. Sel. Top. Signal Process. IEEE 7 (6) (2013) 1112–1123.
[6] B. Li, H. Li, L. Li, J. Zhang, Rate control by R-lambda model for HEVC, Document: JCTVC-K0103, Joint Collaborative Team on Video Coding.
[7] G.J. Sullivan, T. Wiegand, Rate-distortion optimization for video compression, IEEE Signal Process. Mag. 15 (6) (1998) 74–90.
[8] J. Lee, T. Ebrahimi, Perceptual video compression: a survey, Signal Processing, IEEE 6 (6) (Oct. 2012) 684–697.
[9] X. Yang, W. Lin, Z. Lu, X. Lin, S. Rahardja, E. Ong, S. Yao, Rate control for videophone using local perceptual cues, IEEE Trans. Circuits Syst. Video Technol. 15 (4) (2005) 496–507.
[10] Y. Liu, Z.G. Li, Y.C. Soh, Region-of-interest based resource allocation for conversational video communication of H.264/AVC, IEEE Trans. Circuits Syst. Video Technol. 18 (1) (2008) 134–139.
[11] Z. Li, S. Qin, L. Itti, Visual attention guided bit allocation in video compression, Image Vis. Comput. 29 (1) (2011) 1–14.
[12] M. Xu, X. Deng, S. Li, Z. Wang, Region-of-interest based conversational HEVC coding with hierarchical perception model of face, IEEE J. Sel. Top. Signal Process 8 (3) (2014) 475–489.
[13] L. Xu, D. Zhao, X. Ji, L. Deng, S. Kwong, W. Gao, Window-level rate control for smooth picture quality and smooth buffer occupancy, IEEE Trans. Image Process. 20 (3) (2011) 723–734.
[14] L. Xu, S. Li, K.N. Ngan, L. Ma, Consistent visual quality control in video coding, IEEE Trans. Circuits Syst. Video Technol. 23 (6) (2013) 975–989.
[15] A. Rehman, Z. Wang, SSIM-inspired perceptual video coding for HEVC, in: IEEE International Conference on Multimedia and Expo (ICME), 2012, pp. 497–502.
[16] Z. Wang, Q. Li, Information content weighting for perceptual image quality assessment, IEEE Trans. Image Process. 20 (5) (2011) 1185–1198.
[17] W.S. Geisler, 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, vol. 3299, 1998, pp. 294–305.
[18] M.G. Martini, C.T. Hewage, Flexible macroblock ordering for context-aware ultrasound video transmission over mobile WIMAX, Int. J. Telemed. Appl. 2010 (2010) 6.
[19] L. Itti, Automatic foveation for video compression using a neurobiological model of visual attention, IEEE Trans. Image Process. 13 (10) (2004) 1304–1318.
[20] M.-C. Chi, M.-J. Chen, C.-H. Yeh, J.-A. Jhu, Region-of-interest video coding based on rate and distortion variations for H. 263+, Signal Process.: Image Commun. 23 (2) (2008) 127–142.
[21] M. Cerf, J. Harel, W. Einhäuser, C. Koch, Predicting human gaze using low-level saliency combined with face detection, Adv. Neural Inf. Process. Syst. 20 (2008) 241–248.
[22] N. Doulamis, A. Doulamis, D. Kalogeras, S. Kollias, Low bit-rate coding of image sequences using adaptive regions of interest, IEEE Trans. Circuits Syst. Video Technol. 8 (8) (1998) 928–934.
[23] D.M. Saxe, R.A. Foulds, Robust region of interest coding for improved sign language telecommunication, IEEE Trans. Inf. Technol. Biomed. 6 (4) (2002) 310–316.
[24] Y. Sun, I. Ahmad, D. Li, Y.-Q. Zhang, Region-based rate control and bit allocation for wireless video transmission, IEEE Trans. Multimed. 8 (1) (2006) 1–10.
[25] M.-C. Chi, C.-H. Yeh, M.-J. Chen, Robust region-of-interest determination based on user attention model through visual rhythm analysis, IEEE Trans. Circuits Syst. Video Technol. 19 (7) (2009) 1025–1038.
[26] A. Cavallaro, O. Steiger, T. Ebrahimi, Semantic video analysis for adaptive content delivery and automatic description, IEEE Trans. Circuits Syst. Video Technol. 15 (10) (2005) 1200–1209.
[27] G. Boccignone, A. Marcelli, P. Napoletano, G. di Fiore, G. Iacovoni, S. Morsa, Bayesian integration of face and low-level cues for foveated video coding, IEEE Trans. Circuits Syst. Video Technol. 18 (12) (2008) 1727–1740.
[28] L.S. Karlsson, M. Sjostrom, Improved ROI video coding using variable Gaussian pre-filters and variance in intensity, in: IEEE International Conference on Image Processing, 2005, ICIP 2005, vol. 2, 2005, pp. 313–316.
[29] D. Chai, K.N. Ngan, Face segmentation using skin-color map in videophone applications, IEEE Trans. Circuits Syst. Video Technol. 9 (4) (1999) 551–564.
[30] M. Wang, T. Zhang, C. Liu, S. Goto, Region-of-interest based dynamical parameter allocation for H.264/AVC encoder, in: IEEE Picture Coding Symposium, 2009, PCS 2009, 2009, pp. 1–4.
[31] Q. Chen, G. Zhai, X. Yang, W. Zhang, Application of scalable visual sensitivity profile in image and video coding, in: IEEE International Symposium on Circuits and Systems, 2008, ISCAS 2008, 2008, pp. 268–271.
[32] B. Li, H. Li, L. Li, J. Zhang, λ domain based rate control for high efficiency video coding, IEEE Trans. Image Process. 23 (9) (2014) 3841–3854.
[33] D. Liu, X. Sun, F. Wu, S. Li, Y.-Q. Zhang, Image compression with edge-based inpainting, IEEE Trans. Circuits Syst. Video Technol. 17 (10) (2007) 1273–1287.
[34] H. Xiong, Y. Xu, Y.F. Zheng, C.W. Chen, Priority belief propagation-based inpainting prediction with tensor voting projected structure in video compression, IEEE Trans. Circuits Syst. Video Technol. 21 (8) (2011) 1115–1129.
[35] L. Cheng, S. Vishwanathan, Learning to compress images and videos, in: Proceedings of the 24th International Conference on Machine Learning, ACM, 2007, pp. 161–168.
[36] X. He, M. Ji, H. Bao, A unified active and semi-supervised learning framework for image compression, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 65–72.
[37] A. Levin, D. Lischinski, Y. Weiss, Colorization using optimization, in: ACM Transactions on Graphics (TOG), vol. 23, ACM, 2004, pp. 689–694.
[38] E. Kavitha, M.A. Ahmed, A machine learning approach to image compression, Int. J. Technol. Comput. Sci. Eng. 1 (2) (2014) 70–81.
[39] M. Xu, S. Li, J. Lu, W. Zhu, Compressibility constrained sparse representation with learnt dictionary for low bit-rate image compression, IEEE Trans. Circuits Syst. Video Technol. 24 (10) (2014) 1743–1757.
[40] Y. Sun, M. Xu, X. Tao, J. Lu, Online dictionary learning based intra-frame video coding, Wirel. Pers. Commun. 74 (4) (2014) 1281–1295.
[41] S. Mallat, F. Falzon, Analysis of low bit rate image transform coding, IEEE Trans. Signal Process. 46 (4) (1998) 1027–1042.
[42] X.W. Marta Karczewicz, Intra frame rate control based on SATD, Document: JCTVC-M0257, Joint Collaborative Team on Video Coding.
[43] B. Li, H. Li, L. Li., Adaptive bit allocation for R-λ model rate control in HM, Document: JCTVC-M0036, Joint Collaborative Team on Video Coding.
[44] B. Li, D. Zhang, H. Li, J. Xu, Qp determination by λ value, Document: JCTVC-I0426, Joint Collaborative Team on Video Coding.
[45] B. Li, L. Li, J. Zhang, J. Xu, H. Li, Encoding with fixed lagrange multipliers, Document: JCTVC-I0242, Joint Collaborative Team on Video Coding.
[46] J.M. Saragihand, S.S. Lucey, J.F. Cohn, Face alignment through subspace constrained mean-shifts, in: Proceeding of ICCV, 2009, pp. 1034–1041.
[47] S. Ren, X. Cao, Y. Wei, J. Sun, Face alignment at 3000 fps via regressing local binary features, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1685–1692.
[48] B.A.C.L.K. Seshadrinathan, R. Soundararajan, Study of subjective and objective quality assessment of video, IEEE Trans. Image Process. 19 (6) (2010) 1427–1441.