Subjective Rate-distortion Optimization in HEVC with Perceptual Model of Multiple Faces

Yufan Liu #, Haoji Hu *, Mai Xu #

# School of Electronic and Information Engineering, Beihang University, Beijing, 100191, China
** College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, 310027, China*

Corresponding Author: Mai Xu(maixu@buaa.edu.cn)

***Abstract*—This paper proposes a novel perceptual video coding approach with a perceptual model of multiple faces, to improve the coding efficiency of HEVC in video conferencing scenarios. For the perceptual model, a latest active appearance model (AAM) is used to detect multiple faces in a video frame. Then, the perceptual model of multiple faces can be established on the basis of the detected multiple faces. With the established perceptual model of multiple faces, all faces in a video frame can be taken into account for subjective rate-distortion optimization, which is based on the state-of-the-art $r-\lambda$ rate control scheme of HEVC. As such, the perceptual video coding can be achieved for HEVC of video conferencing scenarios. Finally, the experimental results validate the effectiveness of the proposed perceptual video coding approach, in terms of subjective quality.**

***Index Terms*—HEVC, perceptual video coding, video conferencing**

I. Introduction

High efficiency video coding (HEVC) [1], as the next generation video coding standard, was formally finalized and approved in 2013. It enables the double compression efficiency over the preceding H.264/AVC standard. However, the amount of video data traffic going over wireless networks is expected to grow 40-fold over the next five years. Thus, even the latest HEVC standard cannot support ever-increasing requirements on transmitting the huge amount of video data.

Fortunately, there exists much perceptual redundancy in HEVC, as humans beings normally focus on a small region of fixation [2], called region-of-interest (ROI) region. Perceptual video coding [3] aims at reducing such perceptual redundancy, via improving the visual quality in ROI regions at the expense of quality degradation in other regions. It thus offers an efficient solution towards the transmission of the huge amount of video data. Indeed, face is an obvious top down cue [4], which attracts extensive human visual attention. Therefore, the past decade has witnessed the exposition of perceptual video coding [5], [6], [7], [8], [9] on video conferencing applications. For H.264/AVC, Liu *et al.* [8] proposed to optimize a linear rate-quantization (R-Q) model, which is based on the MB-level saliency map of each video frame. As such, the ROI region of face can be taken into account in the saliency map to reduce the perceptual redundancy in video conferencing coding. Most recently, Xu et al. [9] proposed to improve the visual quality and meanwhile to reduce the encoding complexity for the latest HEVC standard, via considering the visual attention on ROIs (e.g., face and facial features). However, all above approaches work on the video conferencing scenarios with only one face.

In this paper, we propose a perceptual video coding approach for HEVC, with a novel perceptual model of multiple faces. Specifically, we develop a new perceptual model, which takes into account multiple faces in a video frame. The previous work of perceptual video coding approach for HEVC is based on the rate control model of unified rate quantization (URQ) [10], which is far from the state-of-the-art r-λ rate control scheme [11]. In this paper, we thus propose to optimize the subjective video quality based on the r-λ rate control scheme for HEVC, in which the subjective quality is determined by our proposed perceptual model of multiple faces.

II. Perceptual model of multiple faces

Our face model is based on detecting facial landmarks. The face region, as well as its feature regions (regions of eyes, nose and mouth) are extracted by using a latest active appearance model (AAM) as proposed in [12]. This method builds a unified model for face detection, pose estimation, and landmark estimation, which is based on mixture of trees with a shared part and a series of face models with different rotation angles. Every facial landmark is modeled as a part and global mixtures are used to capture topological changes due to viewpoint. Fig. 1(a) shows the detected facial landmarks of Frame 58, 84 and 106 in video FourPeople.

A. Refinement of Facial Landmarks by Smoothing

Although the proposed method in [12] can correctly detect the facial landmarks for most frames, there still remains situations such as occlusion, big lighting and pose variations in which the proposed method cannot handle. For example, in the middle image of Fig. 1(a), the proposed method lost

978-1-4673-7314-2/15/$31.00 ©2015 IEEE

tracking of facial landmarks for the left and right persons when part of their faces were occluded. In addition, noise and complex backgrounds may also cause detection inaccuracy. Fortunately, the information that the movements of facial landmarks are continuous in videos can be used to tune the detection results. In this paper, we propose a smoothing algorithm to further refine facial landmark detection in videos which utilizes the continuity of facial landmark movements. Suppose $X_{i_1}, X_{i_2}, ..., X_{i_N}$ are position vectors of facial landmarks in Frame $i_1, i_2, ..., i_N$, respectively ($i_1 < i_2 <, ..., < i_N$). Firstly, we do a linear interpolation to frames in which detection is missing. For example, for frame i ($i_{j-1} < i < i_j$), its vector X_i is obtained by

$$X_i = \lambda X_{i_{j-1}} + (1-\lambda) X_{i_j}, \text{ where } \lambda = \frac{i_j - i}{i_j - i_{j-1}}. \quad (1)$$

By the above procedure, we obtain the position vectors $X_1, X_2, ..., X_N$ for all frames. Then, a smoothing algorithm is used to obtain the refined position $X'_1, X'_2, ..., X'_N$. Our goal is to minimize the following function

$$E(X') = \sum_{i=1}^{N} (X'_i - X_i)^2 + C \sum_{i=2}^{N} (X'_i - X'_{i-1})^2, \quad (2)$$

The first term of (2) is the data term, which emphasizes that the smoothed vectors should be similar to its original counterparts. The second term is the smoothness term which penalizes intensive movements of facial landmarks in adjacent frames. By making $\frac{\partial E(X')}{\partial X'} = 0$, we can obtain

$$\begin{bmatrix} X'_1 \\ X'_2 \\ \vdots \\ X'_N \end{bmatrix} = A^{-1} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \quad (3)$$

where A is a Laplacian operator defined as

$$A = \begin{bmatrix} 1+C & -C & 0 & \cdots & 0 & 0 \\ -C & 1+2C & -C & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -C & 1+C \end{bmatrix}, \quad (4)$$

where we set $C = 5$ for all experiments in this paper. Fig. 1 illustrates the landmark detection of three frames (No. 58, 84 and 106) before and after smoothing. It can be seen that the smoothing technique has obtained better performance. For example, in Frame 84, smoothing has generated facial landmarks for the left and right persons while the original method completely lost tracking of their faces. For Frame 106, facial landmarks became more accurate after smoothing, especially for the second person from the right.

B. Weight Assignment

The detected facial landmarks have separated a video frame into different regions. The task of our compression approach is to assign different weighting parameters based on the saliency map of these regions. Assume that there are P faces detected in a video frame. We use $\mathbf{F}_1, \mathbf{F}_2, ..., \mathbf{F}_P$ to represent the P face regions. We further denote the region $\mathbf{B}$ as the background. For each face region $\mathbf{F}_p$, we use $\mathbf{E}_p$, $\mathbf{N}_p$ and $\mathbf{M}_p$ to represent its eye, nose and mouth regions, respectively. The weight ω_i for the i-th pixel is defined as follows,

$$\omega_i = \begin{cases} \omega_B & , \text{if } i \in \mathbf{B} \\ \omega_F & , \text{if } i \in \bigcup_{p=1}^{P} (\mathbf{F}_p - \mathbf{E}_p - \mathbf{N}_p - \mathbf{M}_p) \\ \omega_E & , \text{if } i \in \bigcup_{p=1}^{P} \mathbf{E}_p \\ \omega_N & , \text{if } i \in \bigcup_{p=1}^{P} \mathbf{N}_p \\ \omega_M & , \text{if } i \in \bigcup_{p=1}^{P} \mathbf{M}_p \end{cases}, \quad (5)$$

where ω_B, ω_F, ω_E, ω_N and ω_M are predefined weighting parameters [1] for the background, face (excluding the facial features), eye, nose and mouth regions, respectively. In this paper, we set $\omega_B = 1$, $\omega_F = \omega_N = 20$ and $\omega_E = \omega_M = 30$ for the following experiments.

III. Subjective Rate-Distortion Optimization

For rate control in HEVC, the r-λ rate control scheme [11] minimizes overall distortion D with the constraint on a given target bits R as follows,

$$\min_{\{r_t\}_{t=1}^{M}} D = \Sigma_{t=1}^{T} d_t \qquad \text{s.t. } \Sigma_{t=1}^{T} r_t \leq R, \quad (6)$$

where d_t and r_t are the distortion and target bit for the t-th coding tree unit (CTU). T is the total number of CTUs in the frame. Then, with a Lagrange multiplier λ, (6) can be converted into the following unconstrained optimization problem:

$$\min_{\{r_t\}_{t=1}^{T}} \Sigma_{t=1}^{T} \left(d_t(r_t) + \lambda r_t \right). \quad (7)$$

To solve (7), Li. *et al.* [11] found out that the Hyperbolic model performs better than other models. Such a model can be expressed by $d_t = c_t {r_t}^{-k_t}$, where C_t and k_t are fitting parameters of $d_t(r_t)$, related to the content of the t-th CTU.

For subjective rate-distortion optimization, there exists

$$r_t = \sum_{i \in \mathbf{I}_t} \mathrm{bpw_i}, \quad (8)$$

for each CTU. Here, $\mathbf{I}_t$ is set of pixels related to the t-th CTU. In addition, $\mathrm{bpw_i}$ is the bit per weight (bpw) of the i-th pixel in a video frame. It is calculated according to the estimated ω_i of our perceptual model (Section II). For more details about the calculation on $\mathrm{bpw_i}$, refer to [9]. Then, the averaged bpw in each video frame can be computed by

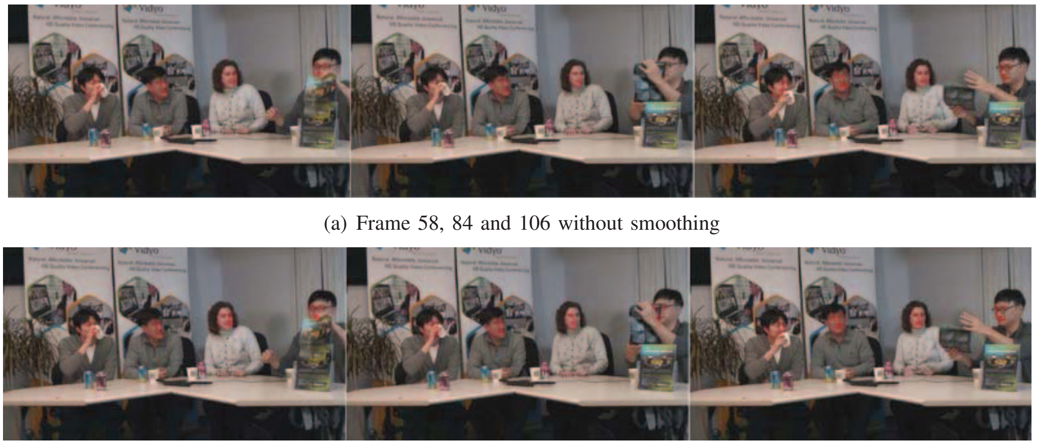
$$\overline{\mathrm{bpw_t}} = \frac{r_t}{|\mathbf{I}_t|}, \quad (9)$$

where $|\mathbf{I}_t|$ stands for the total number of pixels in the t-th CTU. Then, the following subjective r-λ estimation of our approach can be obtained based on (7):

$$\lambda = -\frac{\partial d_t}{\partial r_t} = \alpha_t \cdot \overline{\mathrm{bpw_t}}^{\beta_t}, \quad (10)$$

where α_t and β_t are the parameters for estimating the r-λ relationship. Here, we follow the way of [11] to update these two parameters for each CTU.

[1] We follow the way of [9] to set the values of weights.



(a) Frame 58, 84 and 106 without smoothing

(b) Frame 58, 84 and 106 after smoothing

Fig. 1. The detected facial landmarks before and after the smoothing process.

On the basis of (10), λ can be obtained in our approach. Finally, given the obtained λ, the quantization parameter (QP) of each LCU can be estimated for the subjective rate-distortion optimization. As such, the perceptual model of multiple faces is embedded in the above subjective rate-distortion optimization, for perpetual video coding with weights ω_i.

IV. Experimental results

For experiments, four testing videos are compressed by the conventional HEVC. In this paper, the conventional HEVC approach is implemented by the HM 16.0 software. Then, our approach is embedded into the HM 16.0 software by modifying the $r - \lambda$ rate control scheme as described in Section III.

A. Testing Videos and Parameter Settings

We choose four testing videos 720p: Johnny, KristenAndSara, vidyo1 and FourPeople. There are one to four persons in each of these videos. The frame rate is 60 frame per second (fps) for all videos. In our experiments, the parameter settings are by default which are described in TABLE I. In addition, IPPP structure is also applied for all experiments.

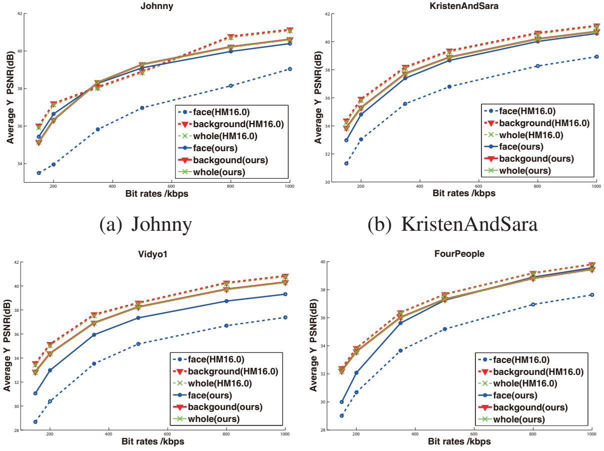
TABLE I
Parameter Settings

Parameter	Setting
GOP Size	4 frames
LCU Size	64*64 pixels
Maximum LCU Depth	3
SAO	Enabled
Search Range of ME	64 pixels

B. Objective Quality Assessment

Fig. 2 illustrates the rate-distortion curves of HEVC and our approach for each video. We individually plot the rate-distortion curves for the face region, the background and the whole image. It can be observed that compared with HEVC, our proposed approach increases the average Y-PSNR of the face regions by approximately 2dB. The average Y-PSNR of the background and the whole region tends to be similar for these two approaches, and HEVC even performs slightly better in some situations. However, since face regions cause much more attention than the background according to HVS, our approach is able to obtain better subjective quality. We further compare the rate-distortion performance in facial feature regions (e.g., eye, nose and mouth regions). Fig.3 shows the rate-distortion curves of HEVC and our approach in facial feature regions (eye, nose and mouth) as well as the whole face. It can be observed that our approach has significant improvement compared with HEVC in facial feature regions. Especially, the quality improvement in the eye and mouth regions reaches to 3dB, more than the average 2dB increase for the whole face region. This is because we have assigned more weights to these regions by our perceptual model.

We also calculate the BD-rate saving of Fig. 2 and Fig.3. Here, the BD-rate is calculated for subjective PSNR of different regions (e.g., PSNR of face region or eye region). For Johnny, our approach can save 59% BD-rate with the same face PSNR and 57% BD-rate with the same eye PSNR. For KirstenAndSara, our approach can save 40% BD-rate with the same face PSNR and 41% BD-rate with the same eye PSNR.



(a) Johnny

(b) KristenAndSara

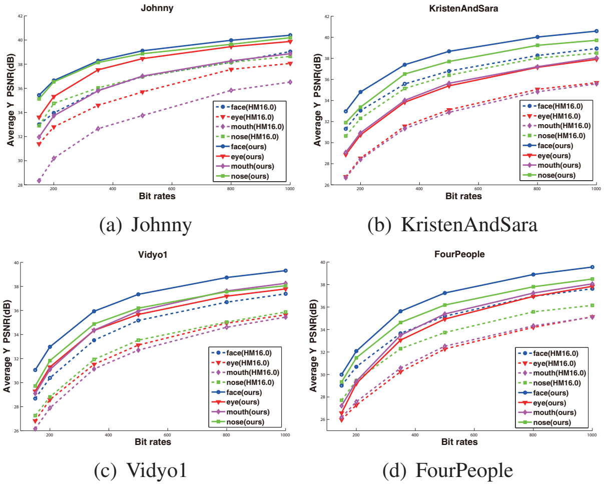
(c) Vidyo1

(d) FourPeople

Fig. 2. Rate-distortion curves of HEVC and our approach in face, background and whole regions of different videos.

TABLE II
DMOS comparison of HEVC and our approach

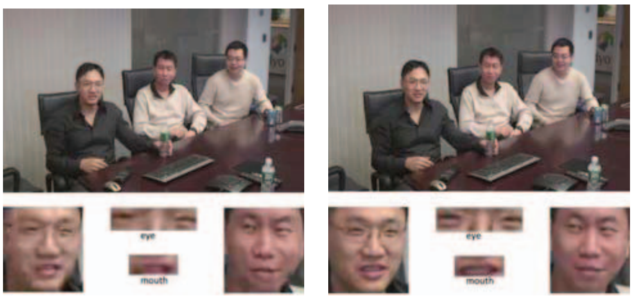
Test video	Bit rate	DMOS HEVC	DMOS ours	DMOS difference
Johnny	150kbps	65.02	43.16	-21.86
	500kbps	40.97	29.53	-11.44
KritenAndSara	150kbps	69.78	52.09	-17.69
	500kbps	50.35	39.40	-10.95
Vidyo1	150kbps	73.14	52.13	-21.04
	500kbps	49.88	33.12	-16.76
FourPeople	150kbps	74.54	54.37	-20.17
	500kbps	51.63	33.48	-18.15



(a) Johnny (b) KristenAndSara

(c) Vidyo1 (d) FourPeople

Fig. 3. Rate-distortion curves of HEVC and our approach in face and facial features regions of different videos.



(a) Conventional HEVC (b) Our approach

Fig. 4. Visual quality comparison of Vidyo1. (a) and (b) are the 49th-frame compressed at 150kbps by HEVC and our approach, respectively.

For Vidyo1, our approach can save 41% BD-rate with the same face PSNR. For FourPeople, our approach can save 36% BD-rate with the same face PSNR. In general, our approach can save bit rates approximately from 35% to 60%.

Fig.4 shows the same frame compressed by HEVC and our approach. Clearly, our approach has better visual quality, especially for the eye and mouth regions.

C. *Subjective Quality Assessment*

We also carry out subjective experiments to evaluate the perceptual quality of both our and HEVC approaches. We choose 12 subjects, 6 males and 6 females aging from 20 to 50, to evaluate these four videos. Firstly, we display several training videos which are labeled as ‘excellent (100-81)’, ‘good (80-61)’, ‘fair (60-41)’, ‘poor (40-21)’ and ‘bad (20-1)’ to train the subjects. Then, videos are shown to them with a random order and the subjects are required to grade each of these videos with score 1 to 100. Each video is compressed by HEVC and our approach at two bit-rates: 150 kbps and 500 kbps. When scores are obtained, we computed the difference mean opinion scores (DMOS), which represent the difference of the compressed video sequences and raw video sequences. Smaller DMOS indicates better subjective quality. TABLE II shows the DMOS of these two approaches. The DMOS values of our approach are much less than those of HEVC, especially at the 150kbps situation. This indicates the superiority of our approach at situations of low bit-rates.

V. Conclusion

In this paper, we have proposed a perceptual video coding approach for HEVC, which optimizes the subjective rate-distortion for video conferencing coding. Specifically, we first proposed a perceptual model of multiple faces for videos. To this end, a method was developed to detect face and facial features of multiple faces in a video fame. Then, different weights are assigned to different regions in a video fame with one or more faces. Based on such weights, the subjective rate-distortion optimization was worked out in our approach to improve the visual quality of ROIs, such as face and facial features. Finally, experimental results verified that our approach enjoys the improvement of our approach over HEVC, in terms of subjective quality.

Acknowledgment

This work was supported by the NSFC projects under Grants 61202400, 61573037, and 61202139, National Key Basic Research Project of China (973 Program 2012CB316400 and 2013CB329006), and Natural Science Foundation of Zhejiang Province (Grant No. LQ12F02014).

References

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

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

[3] J. Lee and T. Ebrahimi, “Perceptual video compression: a survey,” *IEEE Journal of Selected Topics in Signal Processing*, pp. 684–697, 2012.

[4] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, “Predicting human gaze using low-level saliency combined with face detection,” *Advances in neural information processing systems*, vol. 20, 2008.

[5] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, “Low bit-rate coding of image sequences using adaptive regions of interest,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 8, no. 8, pp. 928–934, 1998.

[6] D. M. Saxe and R. A. Foulds, “Robust region of interest coding for improved sign language telecommunication.” *IEEE transactions on information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 6, no. 4, pp. 310–316, 2002.

[7] Y. Sun, I. Ahmad, D. Li, and Y.-Q. Zhang, “Region-based rate control and bit allocation for wireless video transmission,” *Multimedia, IEEE Transactions on*, vol. 8, no. 1, pp. 1–10, 2006.

[8] Y. Liu, Z. G. Li, and Y. C. Soh, “Region-of-interest based resource allocation for conversational video communication of h. 264/avc,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 1, pp. 134–139, 2008.

[9] M. Xu, X. Deng, S. Li, and Z. Wang, “Region-of-interest based conversational hevc coding with hierarchical perception model of face,” *IEEE Journal of Selected Topics on Signal Processing*, vol. 8(4), 2014.

[10] H. Choi, J. Yoo, J. Nam, D. Sim, and I. Bajic, “Pixel-wise unified rate-quantization model for multi-level rate control,” *Journal of Selected Topics in Signal Processing*, 2013.

[11] B. Li, H. Li, L. Li, and J. Zhang, “λ domain based rate control for high efficiency video coding,” *Image Processing, IEEE Transactions on*, vol. 23, no. 9, Sep. 2014.

[12] X. Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in *Computer Vision and Pattern Recognition, 2012. CVPR 2012. IEEE International Conference on*, Providence, RI, 2012, pp. 2879–2886.