SSIM-Inspired Perceptual Video Coding for HEVC

Abdul Rehman and Zhou Wang
Dept. of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada
abdul.rehman@uwaterloo.ca, zhouwang@ieee.org

***Abstract*—Recent advances in video capturing and display technologies, along with the exponentially increasing demand of video services, challenge the video coding research community to design new algorithms able to significantly improve the compression performance of the current H.264/AVC standard. This target is currently gaining evidence with the standardization activities in the High Efficiency Video Coding (HEVC) project. The distortion models used in HEVC are mean squared error (MSE) and sum of absolute difference (SAD). However, they are widely criticized for not correlating well with perceptual image quality. The structural similarity (SSIM) index has been found to be a good indicator of perceived image quality. Meanwhile, it is computationally simple compared with other state-of-the-art perceptual quality measures and has a number of desirable mathematical properties for optimization tasks. We propose a perceptual video coding method to improve upon the current HEVC based on an SSIM-inspired divisive normalization scheme as an attempt to transform the DCT domain frame prediction residuals to a perceptually uniform space before encoding. Based on the residual divisive normalization process, we define a distortion model for mode selection and show that such a divisive normalization strategy largely simplifies the subsequent perceptual rate-distortion optimization procedure. We further adjust the divisive normalization factors based on local content of the video frame. Experiments show that the proposed scheme can achieve significant gain in terms of rate-SSIM performance when compared with HEVC.**

***Index Terms*—SSIM index; HEVC; rate distortion optimization; residual divisive normalization;**

I. Introduction

Over the past years, we have observed an exponential increase in the demand for video services. Recent advances in video capturing and display technologies will increase the presence of high resolution and high quality contents in digital video coding applications. It is therefore expected that both the storage space and bandwidth capacity involved in visual content production, storage, and delivery will be stressed to fulfil the new resolution and quality requirements. This scenario demands for the need to significantly improve the compression performance of the current state-of-the-art H.264/AVC standard. The aforementioned need has gained evidence with the recent activities on high-performance video coding by ISO/IEC Moving Picture Experts Group (MPEG) and the ITU-T Video Coding Experts Group (VCEG) which have joined efforts through the so-called Joint Collaborative Team on Video Coding (JCTVC) to develop a high efficiency video coding (HEVC) standard.

The main objective of a video coding techniques is to optimize the perceptual quality D of the reconstructed video with the number of used bits R subjected to a constraint R_c, which can be expressed by

$$\min\{D\} \quad \text{subject to} \quad R \leq R_c. \tag{1}$$

The desirable distortion model, D, used in the video coding framework should correlate with the perceived distortion of the Human Visual System (HVS), which is the ultimate consumer of the video content. The existing video coding techniques typically use the sum of absolute difference (SAD) or sum of square difference (SSD) as the model for distortion which have been widely criticized in the literature because of their poor correlation with perceptual image quality. Recently, a great deal of effort has been put into the development of advanced quality assessment methods, among which the structural similarity (SSIM) index [1]–[3] achieves a good tradeoff between complexity and quality prediction accuracy, and has become one of the most broadly recognized image/video quality metrics in the past 5 years by both academic researchers and industrial implementers. In recent years, SSIM based video coding techniques have received increasing amount of attention. For example, it has been incorporated into motion estimation, mode selection and rate control [2], [4]–[6].

In this work, we aim to modify the distortion model, D, in (1) by incorporating SSIM into video coding framework using a divisive normalization method. It has already been shown that the main difference between SSIM and MSE is in a locally adaptive divisive normalization process [7]. In general, divisive normalization is recognized as a perceptually and statistically motivated non-linear image representation model [8]. It is shown to be a useful framework that accounts for the masking effect in human visual system, which refers to the reduction of the visibility of an image component in the presence of large neighboring components. It has also been found to be powerful in modeling the neuronal responses in the human perceptual systems [9]. Divisive normalization has been successfully applied in image quality assessment [10], image coding [11], video coding [12] and image denoising [8].

II. SSIM-Inspired Divisive Normalization Scheme for HEVC

Motion compensated inter-prediction plays an important role in the existing hybrid video codec. In this work, we follow this framework, where previously coded frames are used to predict the current frame and only residuals after prediction are coded.

978-0-7695-4711-4/12 $26.00 © 2012 IEEE
DOI 10.1109/ICME.2012.175

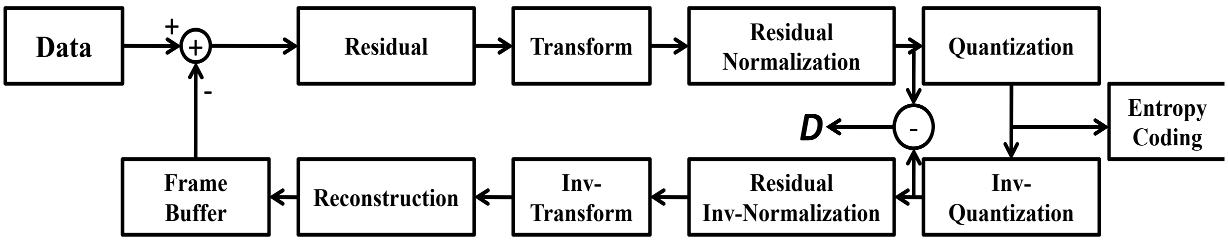


Fig. 1. Framework of the proposed scheme

A. Divisive Normalization Scheme

Assuming $C(k)$ to be the k^{th} DCT transform coefficient of a residual block, then the normalized coefficient is computed as $C(k)' = C(k)/f$, where f is a positive normalization factor which is calculated as the energy of a cluster of neighboring coefficients.

The quantization process of the normalized residuals for a given predefined Q_s can be formulated as

$$\begin{aligned} Q(k) &= sign\{C(k)'\}round\{\frac{|C(k)'|}{Q_s} + p\} \\ &= sign\{C(k)\}round\{\frac{|C(k)|}{Q_s \cdot f} + p\} \end{aligned} \tag{2}$$

where p is the rounding offset in the quantization.

At the decoder, the de-quantization and reconstruction of $C(k)$ is performed as follows

$$\begin{aligned} R(k) &= R(k)' \cdot f = Q(k) \cdot Q_s \cdot f \\ &= sign\{C(k)\}round\{\frac{|C(k)|}{Q_s \cdot f} + p\} \cdot Q_s \cdot f \end{aligned} \tag{3}$$

The purpose of the divisive normalization process is to convert the transform residuals into an perceptually uniform space. Thus the factor f determines the perceptual importance of each of the corresponding transform coefficient. The proposed divisive normalization scheme can be interpreted in two ways. An adaptive normalization factor is applied, followed by quantization with a predefined fixed step Q_s. Alternatively, an adaptive quantization matrix is defined for each MB and thus each coefficient is quantized with a different quantization step.

In the context of still image processing and coding, several different approaches have been used to derive the normalization factor, which can be defined as the sum of the squared neighboring coefficients plus a constant [11], or derived from a local statistical image model [13]. In this work, our objective is to optimize the SSIM index, therefore, we employ a convenient approach based on the DCT domain SSIM index.

The DCT domain SSIM index was first presented by Channappayya *et al.* [14].

$$\begin{aligned} SSIM(\mathbf{x}, \mathbf{y}) = &\{1 - \frac{(X(0) - Y(0))^2}{X(0)^2 + Y(0)^2 + N \cdot C_1}\} \times \\ &\{1 - \frac{\frac{\sum_{k=1}^{N-1}(X(k)-Y(k))^2}{N-1}}{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1} + C_2}\} \end{aligned} \tag{5}$$

where $X(k)$ and $Y(k)$ represent the DCT coefficients for the input signals $\mathbf{x}$ and $\mathbf{y}$, respectively. C_1 and C_2 are used to avoid instability when the means and variances are close to zero and N denotes the block size. This equation implies that the SSIM index is composed of the product of two terms, which are the normalized squared errors of DC and AC coefficientss. Moreover, the normalization is conceptually consistent with the light adaptation (also called luminance masking) and contrast masking effect of HVS.

The HEVC codec uses square-shaped coding tree block (CTB) as a basic unit that may have various sizes. All processing except frame-based loop filtering is performed on a CTB basis, including intra/inter prediction, transform, quantization and entropy coding. In HEVC, coupled with CTB, a basic unit for the prediction mode is the prediction unit (PU), which may be of various sizes and is not necessarily rectangular. In addition to the CTB and PU definitions, the transform unit (TU) for transform and quantization is defined separately in HEVC. The size of TU may be as large as the size of the CTB. TU is always square and is constrained to the range between 4×4 and 64×64.

Since the local statistics do not change significantly within each TU, we divide each TU into l sub-TUs for DCT transform and use $X_i(k)$ to indicate the $k-th$ DCT coefficient in the $i-th$ sub-TU. As the SSIM index differentiates between the DC and AC coefficients, we use separate normalization factors for AC and DC coefficients, respectively. The normalization factors for DC and AC coefficients in each TU are defined as

$$f_{dc} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{X_i(0)^2 + Y_i(0)^2 + N \cdot C_1}}{\mathbb{E}(\sqrt{X(0)^2 + Y(0)^2 + N \cdot C_1})} \tag{6}$$

$$f_{ac} = \frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(X_i(k)^2+Y_i(k)^2)}{N-1} + C_2}}{\mathbb{E}(\sqrt{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1} + C_2})} \tag{7}$$

where $\mathbb{E}(\cdot)$ denotes the mathematical expectation operator over all TUs in the whole frame. The denominator determines relative perceptual importance of each TU. A higher value of the normalization factor, f, of a TU implies that it has relatively lower perceptual importance and can bear more distortion in MSE sense for the same perceptual quality. The proposed divisive normalization scheme aims to achieve uniform perceptual quality over the whole frame by taking the

$$
\begin{aligned}
SSIM(\mathbf{x},\mathbf{y}) &= \{1-\frac{(C(0)'\cdot f_{dc}-R(0)'\cdot f_{dc})^2}{X(0)^2+Y(0)^2+N\cdot C_1}\}\times\{1-\frac{\frac{\sum_{k=1}^{N-1}(C(k)'\cdot f_{ac}-R(k)'\cdot f_{ac})^2}{N-1}}{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1}+C_2}\} \\
&\approx \{1-\frac{(C(0)'-R(0)')^2}{\mathbb{E}(\sqrt{X(0)^2+Y(0)^2+N\cdot C_1})^2}\}\times\{1-\frac{\frac{\sum_{k=1}^{N-1}(C(k)'-R(k)')^2}{N-1}}{\mathbb{E}(\sqrt{\frac{\sum_{k=1}^{N-1}(X(k)^2+Y(k)^2)}{N-1}}+C_2)^2}\}
\end{aligned}
\tag{4}
$$

quality of the TU with energy equal to the expected energy value, over the whole frame, as the reference for the target quality. Consequently, a value of f higher than 1 results in lower number of bits and vice versa.

As a result of the use of f_{dc} and f_{ac}, the SSIM index in the divisive normalization framework can be expressed as in (4), which implies that in the divisive normalized space, the SSIM index is independent of the reference signals and all the TUs can be treated as perceptually identical. As the clearly visible distortion regions will be more apparent from the human visual point of view [15], transforming all the coefficients into the perceptual uniform domain is also a convenient approach to improve the perceptual quality according to the philosophy behind distortion-based pooling [16].

In video coding, these normalization factors need to be computed at both the encoder and the decoder. However, before coding the current frame, the distorted TUs are not available, which creates a chicken or egg causality dilemma. Moreover, at the decoder side, the original TU is not accessible either. Therefore, the normalization factors defined in (6) and (7) cannot be directly applied in this framework. To overcome this problem, we propose the use of predicted TU for the calculation of the normalization factors as it is available at both the encoder and the decoder. In this way, we do not need to transmit any additional overhead information to the decoder. As a result, we can approximate the normalization factor using

$$
f'_{dc}=\frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{2Z_i(0)^2+N\cdot C_1}}{\mathbb{E}(\sqrt{2Z(0)^2+N\cdot C_1})} \tag{8}
$$

$$
f'_{ac}=\frac{\frac{1}{l}\sum_{i=1}^{l}\sqrt{\frac{\sum_{k=1}^{N-1}(Z_i(k)^2+s\cdot Z_i(k)^2)}{N-1}+C_2}}{\mathbb{E}(\sqrt{\frac{\sum_{k=1}^{N-1}(Z(k)^2+s\cdot Z(k)^2)}{N-1}+C_2})} \tag{9}
$$

where $Z_i(k)$ is the $k-th$ DCT coefficient of the $i-th$ prediction sub-TU predicted pixels for each mode (all inter and intra modes) used in the rate distortion optimization process.

In order to compensate for the loss of AC energy, we use a factor s to bridge the difference between the energy of AC coefficients in the prediction TU and the original TU, which can be defined as

$$
s=\frac{\mathbb{E}(\sum_{k=1}^{N-1}X(k)^2)}{\mathbb{E}(\sum_{k=1}^{N-1}Z(k)^2)}. \tag{10}
$$

In [12] it has been shown that s exhibits an approximately linear relationship with Q_s, which can be modeled empirically as

$$
s=1+0.005\cdot Q_s. \tag{11}
$$

The divisive normalization factor is spatially adaptive and depends on the content of the TU and determines the relative perceptual importance of each TU. The TUs which are less important are quantized coarsely with respect to the more important TUs. The expected values of DC and AC energies are used as the reference point to determine the importance of each TU. The TUs with higher energy value than the mean energy value, over the whole frame, have effectively higher QP values than that of the frame and the TUs with lower energy value, have effectively lower QP values. By doing so, we are borrowing bits from the regions which are perceptually less important and using them for the regions with more perceptual relevance, as far as SSIM is concerned, such that all the regions in the frame conceptually have the same perceptual distortion.

It is important to note that the reference point, mean AC and DC energy values, is highly dependent on the content of the video frame. The frames with significant texture regions have high mean AC and DC energy values and are likely to achieve more perceptual improvement for the same rate as compared to the frames with less texture regions as there are many potential candidates with high energy to take bits from and also many areas with moderate energy to give bits to. Subsequently, we perform adjustment of the divisive normalization factors based on the local content of the video frame and found that such content-based adjustment of divisive normalization factors is helpful in improving the robustness of the performance gain across different contents. The video content can be characterized by a local complexity measure computed as local contrast, local energy or local signal activities. We characterize the local complexity by the standard deviation of the energy values of the local 4×4 blocks. A histogram is created to examine the distribution of the DC and AC energy values. The normalization factors for the local blocks with very large or very small energy values is limited to a maximum or minimum value respectively which is determined based on standard deviation of the histogram.

B. Perceptual Rate Distortion Optimization for Mode Selection

The RDO process in video coding can be expressed by minimizing the perceived distortion D with the number of used bits R subjected to a constraint R_c, which can be converted to an unconstrained optimization problem by

$$
\min\{J\} \quad \text{where} \quad J=D+\lambda\cdot R \tag{12}
$$

where J is called the Rate Distortion (RD) cost and λ is known as the Lagrange multiplier which controls the trade-off between R and D.

In the conventional RDO scheme, distortion models such as SAD and SSD are used in actual implementations, but they are widely criticized for not exhibiting good correlation with perceived quality. Here we use a new distortion model that is consistent with the residual normalization process. As illustrated in Fig. 1, for each TU, the distortion model is defined as the SSD between the normalized DCT coefficients, which is expressed by

$$\begin{aligned} D &= \sum_{i=1}^{l} \sum_{k=0}^{N-1} (C_i(k)' - R_i(k)')^2 \\ &= \sum_{i=1}^{l} \frac{(X_i(0) - Y_i(0))^2}{f'^2_{dc}} + \frac{\sum_{N=1}^{N-1} (X_i(k) - Y_i(k))^2}{f'^2_{ac}} \end{aligned} \tag{13}$$

Based on (12), the RDO problem is given by

$$\min\{J\} \text{ where } \quad J = \sum_{i=1}^{l} \sum_{k=0}^{N-1} (C_i(k)' - R_i(k)')^2 + \lambda_{HEVC} \cdot R \tag{14}$$

where λ_{HEVC} indicates the Lagrange multiplier defined in HEVC coding.

From the residual normalization point of view, the distortion model calculates the SSD between the normalized original and distorted DCT coefficients, as shown in Fig. 1. Therefore, we can still use the Lagrange multiplier defined in HEVC, λ_{HEVC}, in this perceptual RDO scheme.

C. Implementation Issues

It is important to note that the proposed scheme is completely compatible with any frame type supported by HEVC, as well as any size or shape choices of CTB, PU and TU, which create significant complications as opposed to the macroblock (MB) structure defined in previous video coding standards such as H.264/AVC. First, the expected values of local divisive normalization factors (the denominator in (8) and (9)) are obtained by first dividing the predicted current frame into 4×4 blocks (the greatest common divisor size for CTB, PU and TU) and then averaged over the whole frame. This avoids the problem of variable sizes of TU that create uneven number of DCT coefficients, and thus reduces the difficulty in estimating the expected values of the divisive normalization factor. Second, the divisive normalization factor for each 4×4 block is computed in pixel domain rather than DCT transform domain. Since DCT is a unitary transform that obeys Parseval's theorem, we have

$$\mu_x = \frac{\sum_{i=0}^{N-1} x(i)}{N} = \frac{X(0)}{\sqrt{N}}, \tag{15}$$

$$\sigma_x^2 = \frac{\sum_{i=1}^{N-1} X(i)^2}{N-1}, \quad \sigma_{xy} = \frac{\sum_{i=1}^{N-1} X(i)Y(i)}{N-1}. \tag{16}$$

As a result, although our algorithm is derived in DCT domain, it is not necessary to perform actual DCT transform for each block in order to perform residual normalization. It allows us to calculate the energy values in pixel domain instead of DCT domain. Since the pixel values used to calculate the energy values are available at the decoder as well, (15) and (16) can also be employed at the decoder. Third, the divisive normalization factor is spatially adaptive but coincides with individual TU. In other words, every TU is associated with a single set of divisive normalization factors but different from other TUs. The normalization matrix thus varies based on the size of TU. However, only two divisive normalization factors are used, one for the DC coefficient and the other for all AC coefficients. Since each TU may contain multiple 4×4 blocks, the divisive normalization factor for each TU is estimated by averaging the divisive normalization factors computed for all 4×4 blocks contained in the TU.

III. Validations

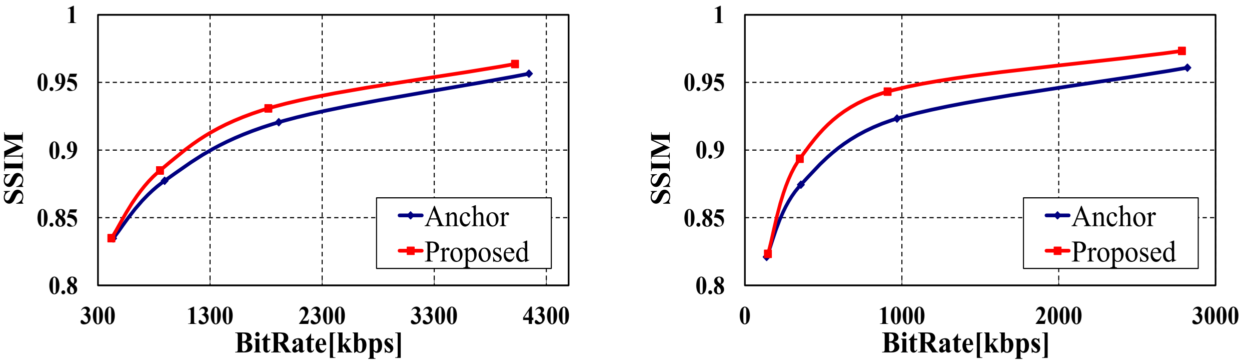
To validate the accuracy and efficiency of the proposed divisive normalization representation based perceptual video coding scheme, we integrated our scheme into the HEVC reference software HM3.0. All test video sequences are in YCbCr 4:2:0 format. We use the standard configuration file for low-delay conditions with IPPP GOP structure and compare our scheme with the HEVC coding schemes in various aspects, including the R-D performance, the coding and decoding complexities and the visual performance. The SSIM index for the whole video sequence are obtained by simply averaging the respective values of individual frames. We employ the method proposed in [17] to calculate the differences between two RD curves which is also used by JCTVC to compare the performance of various algorithms.[1]. The QP values used to obtain the RD curves are 22, 27, 32 and 37, respectively.

Sequence	Resolution	Δ R	Δ SSIM
BasketBallPass	WQVGA	-10.9%	0.008
RaceHorses	WQVGA	-3.0%	0.003
BlowingBubbles	WQVGA	-1.3%	0.001
BQ Square	WQVGA	-30.1%	0.018
PartyScene	WVGA	-2.4%	0.002
BasketBallDrill	WVGA	-16.6%	0.011
Vidyo1	720p	-3.08%	0.002
BQTerrace	1080p	-2.2%	0.002
Average		-8.7%	0.006

TABLE I
PERFORMANCE COMPARISON OF THE PROPOSED SCHEME WITH HEVC

Table I shows the rate savings achieved using proposed scheme for various standard test sequences. It can be observed that over a wide range of test sequences with resolutions from WQVGA to 1080p, our proposed scheme achieves average rate reduction of 8.7% for the same SSIM value and the maximum coding gain is 30.1%. Therefore, the divisive normalization mechanism on average can substantially improve the rate-distortion performance of HEVC. However, the performance

[1] Since R-SSIM curve exhibits a similar shape as R-PSNR curve, we use the same tool proposed in [17] to calculate the average of SSIM differences.



(a) BasketBallDrill

(b) BQSuare

Fig. 2. Rate-SSIM performance comparison between HEVC and the proposed video coding scheme

improvement varies quite significantly, depending on the content of the video frame being encoded. In general, the video frames that have large variations in terms of the texture content often result in more performance gain. Table I also shows the average improvement in terms of SSIM index for the same rate.

The R-D performance for sequences with various resolutions are shown in Fig. 2. In general, the performance gap between the proposed method and the HEVC codec is maximum at the mid-range of QP values. Following are the possible reasons for such a trend. At high bit rate, the quantization step is relatively smaller and thus the differences of quantization steps among the TUs are not significant. At low bit rate, since the AC coefficients are severely distorted, the normalization factors derived from the prediction frame do not precisely represent the properties of the original frame.

When evaluating the coding complexity overhead, we calculate ΔT with

$$\Delta T = \frac{T_{pro} - T_{HEVC}}{T_{HEVC}} \times 100\% \tag{17}$$

where T_{HEVC} and T_{pro} indicate the total coding time for the sequence with the HEVC and the proposed coding schemes, respectively. The average encoding overhead is 7.5% and 9% is the average decoding overhead.

Figure 3 visually compares the proposed scheme with HEVC. For a fair comparison, the bit rate for the proposed scheme is lower than that of HEVC. However, since our proposed divisive normalization scheme is based on SSIM index optimization, higher SSIM and lower PSNR values are achieved. It can be observed by visual comparison of the reconstructed frame with the original frame, the proposed method achieves significantly better visual quality for the same rate. Furthermore, the quality improvement of the reconstructed frame by the proposed scheme is evident from the SSIM maps. The proposed method does a better job in preserving the texture present in the original frame as depicted by the overall brighter SSIM map of the reconstructed frame. It can also be observed that the distortion distribution of the proposed scheme is more uniform across space and more information and details have been preserved. The visual quality improvement is due to the fact that we perform coding algorithms in a perceptual uniform space which can result in a better R-D performance from perceptual point of view.

IV. Conclusion

We proposed an SSIM-inspired novel residual divisive normalization scheme for perceptual video coding. The novelty of the scheme lies in divisively normalizing the transform coefficients based on the DCT domain SSIM index and defining a new distortion model for the subsequent rate distortion optimization. The proposed scheme demonstrates superior performance as compared to the HEVC video codec by offering significant rate reduction, while keeping the same level of SSIM values.

References

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[2] Y. H. Huang, T. S. Ou, P. Y. Su, and H. H. Chen, "Perceptual rate-distortion optimization using structural similarity index as quality metric," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 20, pp. 1614–1624, Nov. 2010.

[3] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication,* special issue on objective video quality metrics, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[4] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Rate-SSIM optimization for video coding," *IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP)*, May. 2011.

[5] H. H. Chen, Y. Huang, P. Su, and T. Ou, "Improving video coding quality by perceptual rate-distortion optimization," *Proc. IEEE Int. Conf. Multimedia Exp*, pp. 1287–1292, Jul. 2010.

[6] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-motivated rate distortion optimization for video coding," *accepted for IEEE Trans. on Circuits and Systems for Video Technology*.

[7] D. Brunet, E. R. Vrscay, and Z. Wang, "Structural similarity-based approximation of signals and images using orthogonal bases," in *Proc. Int. Conf. on Image Analysis and Recognition*, ser. LNCS, vol. 6111, 2010, pp. 11–22.

[8] S. Lyu and E. P. Simoncelli, "Statistically and perceptually motivated nonlinear image representation," *Proc. SPIE Conf. Human Vision Electron. Imaging XII*, vol. 6492, pp. 649 207–1–649 207–15, Jan. 2007.

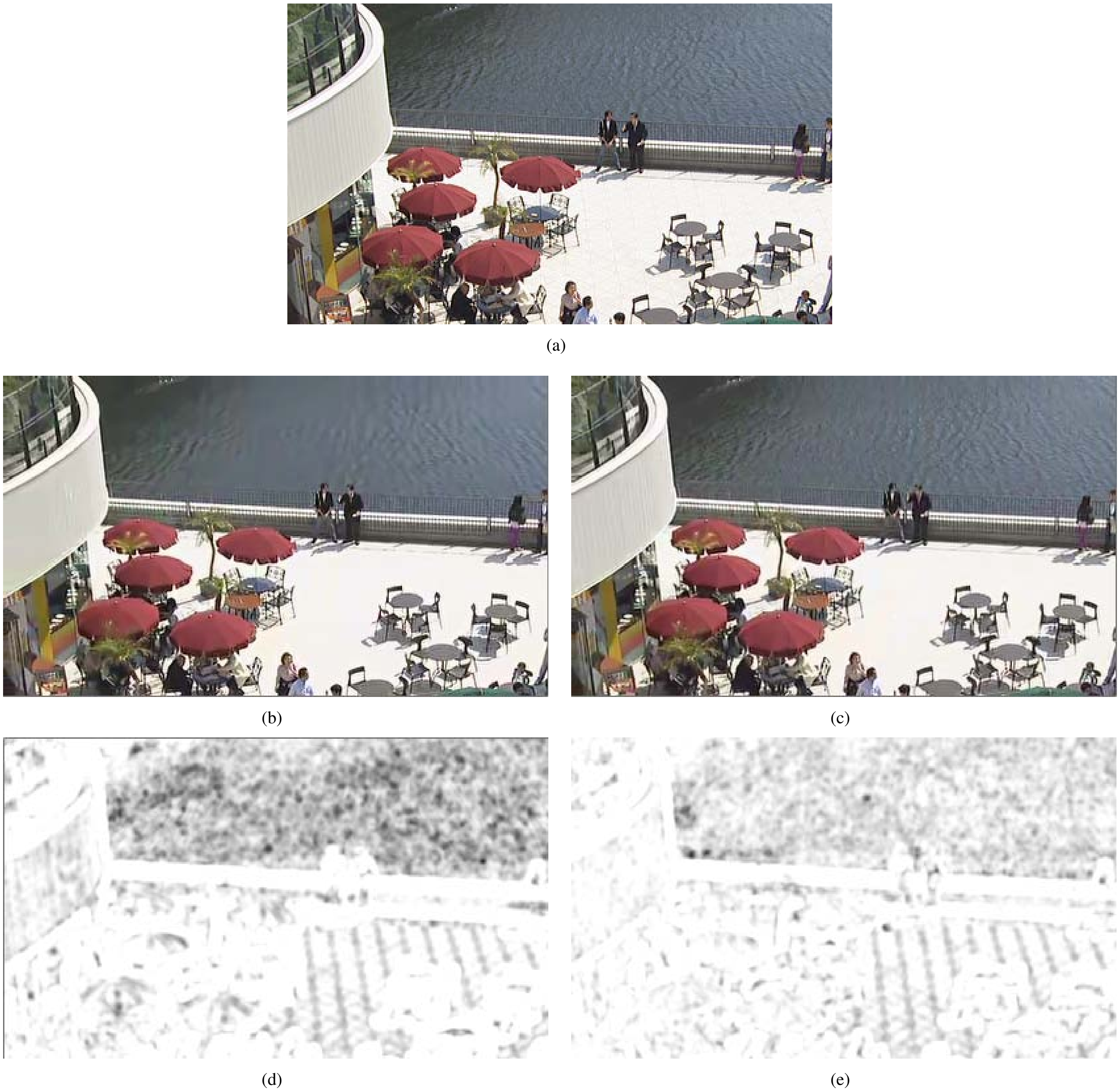


Fig. 3. Visual quality comparison between HEVC and the proposed coding scheme: (a) Original frame; (b) HEVC coded; Bit rate: 356.5192 Kbit/s, SSIM = 0.8744, PSNR = 30.949 dB; (c) Proposed scheme; Bit rate: 349. 6576 Kbit/s, SSIM = 0.8936, PSNR = 29.1254 dB; (d) SSIM map of the HEVC coded video; (e) SSIM map of the video coded using the proposed scheme. In SSIM maps, brighter indicates better quality/larger SSIM value.

[9] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," *Nature Neuroscience*, vol. 4, no. 8, pp. 819–825, August 2001.

[10] A. Rehman and Z. Wang, "Reduced-reference image quality assessment by structural similarity estimation," *accepted by IEEE Trans. on Image Processing*.

[11] J. Malo, I. Epifanio, R. Navarro, and E. P. Simoncelli, "Non-linear image representation for efficient perceptual coding," *IEEE Trans. on Image Processing*, vol. 15, pp. 68–80, Jan. 2006.

[12] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-inspired divisive normalization for perceptual video coding," *IEEE International Conference on Image Processing (ICIP)*, Sep. 2011.

[13] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," *Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 855–861, 2000.

[14] S. Channappayya, A. C. Bovik, and J. R. W. Heathh, "Rate bounds on SSIM index of quantized images," *IEEE Trans. on Image Processing*, vol. 17, pp. 1624–1639, Sep. 2008.

[15] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, 2010.

[16] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. on Image Processing*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[17] G. Bjontegaard, "Calculation of average PSNR difference between RD curves," *Proc. ITU-T Q.6/SG16 VCEG 13th Meeting, Austin, TX*, 2001.