Saliency-Aware Video Compression

Hadi Hadizadeh, *Student Member, IEEE*, and Ivan V. Bajić, *Senior Member, IEEE*

***Abstract*— In region-of-interest (ROI)-based video coding, ROI parts of the frame are encoded with higher quality than non-ROI parts. At low bit rates, such encoding may produce attention-grabbing coding artifacts, which may draw viewer's attention away from ROI, thereby degrading visual quality. In this paper, we present a saliency-aware video compression method for ROI-based video coding. The proposed method aims at reducing salient coding artifacts in non-ROI parts of the frame in order to keep user's attention on ROI. Further, the method allows saliency to increase in high quality parts of the frame, and allows saliency to reduce in non-ROI parts. Experimental results indicate that the proposed method is able to improve visual quality of encoded video relative to conventional rate distortion optimized video coding, as well as two state-of-the art perceptual video coding methods.**

***Index Terms*— Video compression, visual saliency, global motion estimation, DCT.**

I. Introduction

LOSSY image and video encoders are known to produce undesirable compression artifacts at low bit rates [1], [2]. Blocking artifacts are the most common form of compression artifacts in block-based video compression. When coarse quantization is combined with motion-compensated prediction, blocking artifacts propagate into subsequent frames and accumulate, causing structured high-frequency noise or motion-compensated edge artifacts that may not be located at block boundaries, and so cannot be attenuated by deblocking filters that mostly operate on block boundaries [2]. Such visual artifacts may become very severe and attention-grabbing (salient), especially in low-textured regions.

Recently, region-of-interest (ROI) coding of video using computational models of visual attention [3] has been recognized as a promising approach to achieve high-performance video compression [4]–[7]. The idea behind most of these methods is to encode an area around the predicted attention-grabbing (salient) regions with higher quality compared to other less visually important regions. Such a spatial prioritization is supported by the fact that only a small region of 2–5° of visual angle around the center of gaze is perceived with high spatial resolution due to the highly non-uniform distribution of photoreceptors on the human retina [4].

Granting a higher priority to the salient regions, however, may produce visible coding artifacts in areas outside the salient regions where the image quality is lower. Such artifacts may draw viewer's attention away from the naturally salient regions, thereby degrading the perceived visual quality. It is worth pointing out that a visible artifact is not necessarily salient. A particular artifact may be visible if the user is looking directly at it or at its neighborhood, but may go unnoticed if the user is looking elsewhere in the frame. As the severity of the artifact increases, it may become salient and draw user's attention. Hence, a salient artifact is necessarily visible, but a visible artifact need not be salient. Several methods have been developed for detecting visible artifacts, such as the Visible Difference Predictor [8] or Most Apparent Distortion (MAD) [9], but relatively little research has been done on studying artifact saliency and developing methods to reduce it.

In [10], we proposed a saliency-preserving framework for region-of-interest (ROI) video coding, whose main goal was to reduce attention-grabbing coding artifacts in non-ROI parts of the frame in order to keep viewer's attention on ROI parts where the video quality was higher. The method proposed in [10] was based on finding a quantization parameter (QP) matrix for each video frame so that the L_1-norm of the difference between the saliency map of the coded frame and the saliency map of the original raw frame was minimized under a given a target bit rate. In this method, the desired QP matrix was obtained after multiple encodings of each frame, which made the process computationally expensive.

In this paper, we extend our earlier work [10] in four ways. First, we develop a computationally efficient method for saliency estimation in the DCT domain based on the well-known Itti-Koch-Niebur (IKN) saliency model [3], with improved temporal saliency estimation that considers the effect of camera motion. Second, we extend the conventional H.264/AVC rate distortion optimization (RDO) [11] by introducing a saliency distortion term in the distortion metric. Unlike our earlier method [10], in the new method, the saliency of non-ROIs is allowed to decrease, and the saliency of ROIs is allowed to increase so long as the quality within ROIs is good. This allows more flexibility in selecting coding parameters while producing visually pleasing results. Third, the complexity of the new method is significantly lower than that of our earlier method [10], which makes it more amenable for real-time applications. This is a consequence of the fact that saliency estimation is performed by reusing some of the data from the coding process. Fourth, we evaluate the proposed method using several objective quality metrics, as well as an extensive subjective study, and compare its performance to two state-of-the-art perceptual video coding approaches.

The paper is organized as follows. In Section II, we present an overview of the IKN saliency model, the rate distortion optimization in H.264/AVC, and several existing ROI-based video coding methods. The proposed methods are

Manuscript received February 10, 2013; revised June 6, 2013 and August 20, 2013; accepted August 21, 2013. Date of publication September 20, 2013; date of current version November 7, 2013. This work was supported by the NSERC under Grant RGPIN 327249. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joan Serra-Sagrista.

The authors are with the School of Engineering Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada (e-mail: hha54@sfu.ca; ibajic@ensc.sfu.ca).

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

Digital Object Identifier 10.1109/TIP.2013.2282897

描述
改进的结论
模型目的
方法
实现参数
问题
改进
时间
灵活性
复杂度，应用
性能指标
问题
思路
依据
问题
过犹不及
过分关注某点
导致出现相反结果
位置

described in Sections III and IV. Experimental results are presented in Section V, followed by conclusions in Section VI. An implementation of the proposed methods can be found at http://geomm.ensc.sfu.ca/software/SAVC-code.rar.

II. Background

A. The IKN Saliency Model

Among the existing bottom-up computational models of visual attention, the Itti-Koch-Niebur (IKN) model [3] is one of the most well-known and widely used. In this model, the visual saliency is predicted by analyzing the input image through a number of pre-attentive independent feature channels, each locally sensitive to a specific low-level visual attribute, such as local opponent color contrast, intensity contrast, and orientation contrast. More specifically, nine spatial scales are created using dyadic Gaussian pyramids, which progressively low-pass filter and down-sample the input image, yielding an image-size-reduction factor ranging from 1:1 (scale zero) to 1:256 (scale eight) in eight octaves [3].

The contrast in each feature channel is computed using a "center-surround" mechanism, which is implemented as the difference between fine and coarse scales: the center is a pixel at scale $c \in \{2, 3, 4\}$, and the surround is the corresponding pixel at scale $s = c + d$, with $d \in \{3, 4\}$. The across-scale difference between two levels of the pyramid is obtained by interpolation to the finer scale and point-by-point subtraction. The obtained contrast (feature) maps are then combined across scales through a normalization operator to create a "conspicuity map" for each feature channel. The conspicuity maps are resized to level 4, and combined together via the same normalization operator to generate a "master saliency map" whose pixel values predict saliency.

A motion and flicker channel were added to the IKN model in [12] to make it applicable to video. The flicker channel is created by building a Gaussian pyramid on the absolute luminance difference between the current frame and the previous frame. Motion is computed from spatially-shifted differences between intensity pyramids from the current and previous frame [12]. The same center-surround mechanism that is used for the intensity, color, and orientation channels is used for computing the motion and flicker conspicuity maps, which are then combined with spatial conspicuity maps into the final saliency map.

B. Rate-Distortion Optimization (RDO)

The H.264/AVC video coding standard supports various block coding modes such as INTER16 × 16, INTER16 × 8, INTER8 × 16, INTER8 × 8, INTRA16 × 16, INTRA4 × 4, etc. [11]. The rate-distortion optimization (RDO) process proposed in H.264/AVC minimizes the following Lagrangian cost function for coding mode selection of each macroblock (MB) [11], [13]:

$$J(\psi|Q, \lambda_R) = D_{MSE}(\psi|Q) + \lambda_R R(\psi|Q), \quad (1)$$

where Q is the quantization step size, $D_{MSE}(\psi|Q)$ and $R(\psi|Q)$ are, respectively, the Mean Squared Error (MSE) and bit rate for coding the current MB in the coding mode ψ with quantization step size Q, and λ_R is the Lagrange multiplier, which quantifies the trade-off between the rate and distortion [13]. The Lagrangian cost function (1) is minimized for a particular value of λ_R. Hence, λ_R has an important role in achieving optimal RD performance [13], [14]. In the H.264/AVC reference software [15], λ_R is computed as

$$\lambda_R = 0.85 \cdot 2^{\frac{(QP-12)}{3}}, \quad (2)$$

where QP is the quantization parameter. The derivation of (2) was based on empirical results under a "high rate" assumption [13], [14], [16]. Although (2) provides a simple and effective method for finding λ_R, it has two main drawbacks. First, it is solely a function of QP, and so it does not consider any property of the input signal, which means that it cannot adapt to the video content. Second, the high rate assumption may not hold in all cases, for example at low bit rates [16].

In the literature, several methods have been proposed to obtain λ_R adaptively based on the video content when MSE is used as the distortion metric [16]–[19]. Most such methods utilize RD models that are based on the distribution of transformed residuals. In particular, they use RD models that have a closed-form expression so that λ_R can be obtained in closed form. For instance, in [16], a Laplace distribution-based RD model was proposed to derive λ_R for each video frame adaptively based on the statistical properties of the transformed residuals. Several methods have shown that adjusting λ_R on the MB level results in better RD performance than λ_R adjustment on the frame level [19]–[21].

Many of the existing methods for RDO utilize the MSE or Sum of Absolute Differences (SAD) as a distortion metric, and they do not consider perceptual aspects. Recently, a number of RDO schemes have been proposed to consider several perceptual aspects of the Human Visual System (HVS). For instance, the authors in [22] proposed a motion-compensated residue signal pre-processing scheme based on just-noticeable-distortion (JND) profile for video compression. A foveated JND model was utilized in [23] for QP and Lagrange multiplier selection in which both the QP and Lagrange multiplier is adjusted for each MB based on the visual noticeable distortion of the MB. Several methods employed the SSIM metric [24] for video coding and RDO [25]–[28]. In [28], the authors utilized SSIM [24] as the distortion metric within the RDO process. They also presented an adaptive Lagrange multiplier selection scheme based on a novel statistical reduced-reference SSIM model and a source-side information combined rate model. Moreover, they proposed a method to adjust the Lagrange multiplier for each MB based on the motion information content and perceptual uncertainty of visual speed perception. In [29], the authors also employed the SSIM as the distortion metric, and weighted the SSIM distortion using the visual saliency of various MBs, with the idea that the perception of distortions is stronger in more salient regions. In [30], a perceptual video coding framework based on a divisive normalization scheme was proposed. In this framework, a normalization scheme is first employed to transform the DCT domain residuals to a perceptually uniform space based on a DCT domain SSIM index. A MB-level perceptual RDO mode decision is then developed based on the proposed

normalization scheme. In [31], a simple method to incorporate SSIM index into the RDO framework for video coding was proposed. This method is based on expressing the SSIM index in terms of the sum-of-square error (SSE), and scaling the Lagrange multiplier in RDO based on local video statistics such as the variance of the luminance component.

In this paper, we present a saliency-aware video compression method that incorporates a novel saliency distortion term into the conventional RDO process to control how saliency is changed after compression. Unlike the existing perceptual video coding methods that either do not use visual saliency or do not consider what happens to saliency after compression, to our knowledge, our earlier method [10] was the first to consider the saliency change after compression, while the present method goes further to model this change, and subsequently allow the saliency to change in a more controlled manner as will be described in Section IV.

C. ROI-Based Video Coding

In the literature, several ROI-based video coding methods have been proposed [4], [6], [7], [23], [32]. For instance, in [7], a ROI-based bit allocation scheme was proposed for H.264/AVC. In this scheme, which was targeted at conversational video, ROI parts are detected using the direct frame difference and skin-tone information. After detecting ROI parts, several coding parameters including QP, coding modes, the number of reference frames, the accuracy of motion vectors, and the search range for motion estimation, were adaptively adjusted at the MB level according to the relative importance of each MB and a given target bit rate. In [32], a skin color-based face detection method was used to locate the ROI, and the rate in different regions is adapted based on a just-noticeable-difference (JND) model. Reference [23] proposed a foveated JND (FJND) model that incorporates both visual sensitivity and foveation features of the HVS. The proposed FJND model was used in H.264/AVC video coding. For each MB, the quantization parameter was optimized based on the distortion visibility thresholds given by the FJND model. The Lagrange multiplier in RDO was adapted to minimize the MB noticeable distortion. In this method, the ROIs were the predicted fixation points obtained as the peaks in the saliency map of each video frame.

In ROI-based video coding, more bits are typically assigned to ROI parts of the frame compared to non-ROI parts. There are two main aspects that differentiate existing ROI coding methods: (1) the way in which ROI is decided, and (2) the way in which bits are allocated to ROI vs. non-ROI. The proposed method differs from the existing work in the area in both aspects. First, to detect ROI, a novel saliency estimation method is employed. It consists of a convex approximation to the spatial IKN saliency combined with a global motion-compensated temporal saliency estimate. The new saliency estimate is computationally more efficient, yet competitive in terms of accuracy with the IKN model for video [12], which is known to be highly accurate in terms of gaze prediction on video [33]. Second, the bit allocation procedure in the proposed method takes into account the way in which saliency is affected by compression. We propose a model for this process and employ it in the RDO procedure to control the way in which saliency is allowed to change after compression. To our knowledge, such approach has not been employed before in the context of video coding. The only prior work to consider controlling saliency after compression was our earlier work [10], however that method was computationally demanding, requiring execution of the saliency estimator after each QP change, and the goal was to minimize saliency change after compression. In the method proposed here, saliency is allowed to change in a controlled manner, as will be described below.

In this following, we first present a method for saliency estimation in video that is inspired by the IKN model reviewed in Section II-A, yet is simpler to compute, and more accurate in scenes with camera motion. We then describe our proposed saliency-aware video compression method in Section IV.

III. Saliency Estimation in Video

The saliency model employed in this work consists of two parts: the spatial saliency component (Section III-A), which is a convex approximation to the spatial IKN saliency [3], and a temporal saliency component (Section III-B), which incorporates global motion compensation to remove the influence of camera motion on the saliency estimate. The two saliency components are described below.

A. Convex Approximation to Spatial IKN Saliency

As discussed in [10], [34], and [35], the IKN saliency model uses the image content in the normalized frequency range $[\pi/256, \pi/16]$ to construct the saliency map of an image. In [34], we proposed a convex approximation to the spatial IKN saliency, which uses the DCT of a given block $\mathbf{X}$ to estimate the saliency of that block. This is achieved by recapturing the portion of the image signal from the normalized frequency range $[\pi/256, \pi/16]$ at the position of $\mathbf{X}$. Since the process of extracting a block from an image involves windowing and spectral down-sampling, which leads to spectral leakage, some energy from the normalized frequency range $[\pi/256, \pi/16]$ of the original image will be present at other frequencies when examining the spectrum of the block $\mathbf{X}$. Analogously, some energy of the original signal from outside of the normalized frequency range $[\pi/256, \pi/16]$ will leak into this range in the process of block extraction. To address this issue, the following approach was taken in [34].

Consider the original image spectrum in the normalized frequency range $[0-\pi]$. The image signal in the normalized frequency range $[\pi/256, \pi/16]$ is considered the "signal," and the signal in the remaining part of the spectrum, $[0-\pi/256)\cup(\pi/16-\pi]$, is considered "noise." After extracting a block from the image, the spectrum of the block $\mathbf{X}$ will be the sum of the leaked spectra of the signal and the noise due to windowing and spectral down-sampling. To extract signal from noise for the purpose of saliency estimation, the approach in [34] employs a Wiener filter in the DCT domain.To obtain the Wiener filter in the DCT domain for a given image resolution and block size, a deterministic $1/f$ 2-D signal [36] that covers

the frequency band $[\pi/256 - \pi/16]$ is generated at the given image resolution. A block whose size is equal to the block size of interest is then extracted from this signal, and its 2-D DCT is computed. Let us denote the resulting DCT by $\mathbf{Z_S}(i, j)$. Similarly, the 2-D DCT of a block extracted from a deterministic $1/f$ 2-D signal that covers the frequency band $[0 - \pi/256) \cup (\pi/16 - \pi]$ is computed. Let us denote the resulting DCT by $\mathbf{Z_V}(i, j)$. The DCT-domain Wiener filter coefficients are then given by [34]

$$\mathbf{H}(j,l) = \frac{\mathbf{Z_S}^2(j,l)}{\mathbf{Z_S}^2(j,l) + \mathbf{Z_V}^2(j,l)}, \tag{3}$$

where $\mathbf{H}(j,l)$ is the (j,l)-th Wiener filter coefficient.

As discussed in [34] and [37], for a MB $\mathbf{X}$, whose 2-D DCT is $\mathbf{Z_X}$, the (un-normalized) spatial saliency is estimated as the energy of the Wiener-filtered signal $\mathbf{Z}_\mathbf{X}^W$, that is

$$\mathcal{S}_s^u(\mathbf{X}) = \sum_{(j,l)} (\mathbf{Z}_\mathbf{X}^W(j,l))^2 = \sum_{(j,l)} \mathbf{H}^2(j,l)\mathbf{Z_X}^2(j,l), \tag{4}$$

where $\mathbf{Z_X}(j,l)$ is the (j,l)-th 2-D DCT coefficient of $\mathbf{X}$, and $\mathbf{H}(j,l)$ is the corresponding Wiener coefficient. In practice, $\mathbf{H}$ can be pre-computed for a given video resolution and block size. If MB $\mathbf{X}$ has multiple color channels (e.g., YUV), the energy in all channels is added together.

To compute the spatial saliency map of a given video frame, equation (4) is applied to all MBs in the frame. The resulting map is then normalized to the range [0, 1] to obtain the final spatial saliency map of that frame. Hence, the normalized spatial saliency of a block $\mathbf{X}$ in the final saliency map is obtained as

$$\mathcal{S}_s(\mathbf{X}) = \frac{1}{\sum_\mathbf{X} \mathcal{S}_s^u(\mathbf{X})} \mathcal{S}_s^u(\mathbf{X}). \tag{5}$$

As discussed in [37], some coefficients in $\mathbf{H}$ may be zero for a given image resolution and block size. Hence, the un-normalized spatial saliency in (4) is monotonically non-decreasing (but not necessarily increasing) with the total DCT energy in the block. However, because the final saliency is normalized over the whole frame, the saliency value of a block may decrease even if the DCT energy of the block increases, provided the energy in some other blocks in that frame increases even more. For further details and accuracy comparisons between the IKN model and the convex approximation described above, the reader is referred to [34] and [37].

B. Global Motion-Compensated Temporal Saliency

While the spatial component of saliency is borrowed from [34], the temporal component, discussed in this section, is different. It is well-known that object motion is one of the strongest attractors of visual attention [38]–[40]. In many existing computational models of visual attention, the temporal saliency is estimated by measuring the local motion contrast [12]. An object with significant motion with respect to its surroundings would be considered as a strong, attention-grabbing "surprise" to the visual system.

In [33], it was observed that the accuracy of the IKN model for video [12] degrades on scenes with camera motion. When the camera moves, the resulting apparent motion of the background competes with foreground object motion and may confuse the saliency model, leading to lower accuracy. To mitigate this problem, similar to [39] and [40], we remove the camera motion prior to computing temporal saliency. To make the process computationally efficient, we use the previous frame's motion field (which is already computed) as an approximation to the current frame's motion field, and run an efficient compressed-domain global motion estimation algorithm [41], followed by global motion compensation, i.e. subtraction of global motion from the motion field. This way, we obtain one global motion-compensated MV (GMC-MV) per 4×4 block. For each MB $\mathbf{X}$, the average magnitude of all GMC-MVs in it is taken as the motion saliency $\mathcal{S}_m(\mathbf{X})$.

In order to obtain the spatio-temporal saliency of $\mathbf{X}$, we combine the spatial and motion saliency of $\mathbf{X}$ using the coherent-normalization-based fusion method from [42] and [43], as follows

$$\mathcal{S}(\mathbf{X}) = (1-\alpha)\mathcal{S}_s(\mathbf{X}) + \alpha\mathcal{S}_m(\mathbf{X}) + \beta\mathcal{S}_s(\mathbf{X})\mathcal{S}_m(\mathbf{X}), \tag{6}$$

where α and β are positive constants. The quantity $\mathcal{S}(\mathbf{X})$ in (6) is referred to as the global motion-compensated (GMC) saliency of $\mathbf{X}$. The first two terms in (6) allow the spatial and temporal saliency to promote a MB independently. On the other hand, the third term in (6) weighs the spatial saliency value by the temporal saliency value and vice versa. Hence, it is a mutual reinforcement term, which promotes those MBs that are salient both spatially and temporally. As mentioned earlier, it is known that motion cues are one of the strongest attractors of visual attention [38]. Hence, in practice, a larger relative weight for the temporal saliency ($\alpha > 0.5$) is recommended. As discussed in Section V-A, we set $\alpha = 0.9$ and $\beta = 1$ in our experiments. The results in Section V-A indicate that the performance of the proposed saliency estimation method is better than the IKN saliency model for video [12] on sequences containing camera motion.

IV. Saliency-Aware Video Compression

The proposed saliency-aware video compression is based on the following principles.

1) Highly salient regions should end up with higher perceptual quality than less salient regions. This means that quality is directed towards the regions that viewers are likely to look at.
2) The coding should attempt to preserve the saliency of various regions, except in the following two cases.
   - If a region is highly salient, then its saliency is allowed to increase after compression, provided the quality remains sufficiently high. The reasoning here is that we don't mind viewers being even more drawn to high-quality regions in the scene.
   - If a region has low saliency to start with, then its saliency is allowed to decrease after compression. The logic here is that low-saliency regions will end up with lower quality, so the less likely the viewer is to look at such regions, the better.

In the remainder of this section, we present procedures for selecting the quantization parameter (QP), the Lagrange

multipliers, and the optimal coding mode, to satisfy the above principles. For each MB in the frame, the QP is assigned first based on MB's saliency, followed by Lagrange multiplier selection and coding mode decision.

A. Macroblock QP Selection

Let QP_f be the quantization parameter of the current video frame, which is provided by an appropriate frame-level rate control algorithm, e.g. [44]–[46]. Let $\mathcal{S}(\mathbf{X}_i)$ be the saliency of the i-th MB $\mathbf{X}_i$, and $\bar{s}$ be the average saliency of all MBs in the current frame. Following the method from [23], the QP for the i-th MB in the current frame is obtained as

$$\mathrm{QP}_i = round\left(\frac{\mathrm{QP}_f}{\sqrt{w_i}}\right), \tag{7}$$

where w_i is obtained through a sigmoid function

$$w_i = a + \frac{b}{1+\exp(-c(\mathcal{S}(\mathbf{X}_i)-\bar{s})/\bar{s})}, \tag{8}$$

and a, b, and c are constants. In our experiments, similar to [23], we set $a = 0.7$, $b = 0.6$, and $c = 4$.

Note that (7) gives the QP of $\mathbf{X}_i$. In H.264/AVC, the relation between QP and the quantization step size Q is

$$Q = 2^{\mathrm{QP}/6} \cdot \nu(\mathrm{QP} \bmod 6),$$

where $\nu(0) = 0.675$, $\nu(1) = 0.6875$, $\nu(2) = 0.8125$, $\nu(3) = 0.875$, $\nu(4) = 1.0$, and $\nu(5) = 1.125$ [11].

B. RDO Mode Decision

In addition to the conventional rate and distortion terms commonly used in the Lagrangian cost function, we introduce a saliency distortion term $D_{sal}(\psi|Q_i, \mathbf{X}_i)$ in order to obtain the optimal coding mode according to the principles outlined above. For the i-th MB, the proposed cost function is

$$\begin{aligned} &J_i(\psi|Q_i, \lambda_{S_i}, \lambda_{R_i}, \mathbf{X}_i) \\ &= D_{MSE}(\psi|Q_i, \mathbf{X}_i) + \lambda_{S_i} D_{sal}(\psi|Q_i, \mathbf{X}_i) + \lambda_{R_i} R(\psi|Q_i, \mathbf{X}_i), \end{aligned} \tag{9}$$

where λ_{S_i} is the Lagrangian multiplier associated with saliency distortion $D_{sal}(\psi|Q_i, \mathbf{X}_i)$, which is the absolute difference between the saliency of the uncompressed i-th MB and the i-th MB coded using coding mode ψ with quantization step size Q_i, that is,

$$D_{sal}(\psi|Q_i, \mathbf{X}_i) = |\mathcal{S}(\mathbf{X}_i) - \mathcal{S}(\widetilde{\mathbf{X}}_i(\psi|Q_i))| \tag{10}$$

where $\mathbf{X}_i$ is the uncompressed i-th MB and $\widetilde{\mathbf{X}}_i(\psi|Q_i)$ denotes the i-th MB coded using coding mode ψ with quantization step size Q_i.

We note that compression generally does not change the direction or magnitude of motion of various regions, except possibly at extremely low bitrates. We will therefore assume that the change in motion saliency in (6) due to compression is negligible compared to the change in spatial saliency. Hence, using (6), $D_{sal}(\psi|Q_i, \mathbf{X}_i)$ can be approximated as

$$D_{sal}(\psi|Q_i, \mathbf{X}_i) = \mu_i \cdot |\mathcal{S}_s(\mathbf{X}_i) - \mathcal{S}_s(\widetilde{\mathbf{X}}_i(\psi|Q_i))|, \tag{11}$$

where

$$\mu_i = 1 - \alpha + \beta \mathcal{S}_m(\mathbf{X}_i). \tag{12}$$

According to (11) and (12), the saliency distortion for a MB is the spatial saliency distortion weighted by the motion saliency of the MB. Hence, other things being equal, the saliency distortion is expected to be larger in regions where the motion saliency is higher. Note that for the purpose of computing the saliency distortion in (11), we need to compute $\mathcal{S}_s(\mathbf{X}_i)$ and $\mathcal{S}_s(\widetilde{\mathbf{X}}_i(\psi|Q_i))$ using (5). To compute (5) for $\mathcal{S}_s(\widetilde{\mathbf{X}}_i(\psi|Q_i))$, we need to know the after-compression saliency of all blocks in the frame. However, because the after-compression saliency of blocks that have not yet been coded is not known, it is a chicken and egg problem. To mitigate this problem, we use the spatial saliency of raw video in the denominator of (5) for estimating $\mathcal{S}_s(\widetilde{\mathbf{X}}_i(\psi|Q_i))$. By doing so, because the denominators in $\mathcal{S}_s(\mathbf{X}_i)$ and $\mathcal{S}_s(\widetilde{\mathbf{X}}_i(\psi|Q_i))$ are the same, normalized and un-normalized saliency distortion are proportional to each other, so either can be used to compute (11), the difference being only the scale factor. Hence, in our experiments, we use the un-normalized saliency given in (4) for computing the saliency distortion in (11).

According to the principles outlined at the beginning of this section, the saliency of highly salient regions (ROIs) is allowed to increase after compression, if the quality of such regions after compression is good. This condition is characterized by

$$\text{Condition A} = \begin{cases} \mathbf{X}_i \in \text{ROI}, & \text{and} \\ \mathcal{S}_s(\mathbf{X}_i) < \mathcal{S}_s(\widetilde{\mathbf{X}}_i(\psi|Q_i)), & \text{and} \\ D_{MSE}(\psi|Q_i, \mathbf{X}_i) < \delta, \end{cases}$$

where δ is a user-defined threshold. In our experiments, we set $\delta = 21$, which is equivalent to a PSNR of about 35 dB, a level usually associated with reasonably good quality. Also, the saliency of low-salient regions (non-ROIs) is allowed to decrease after compression. Such condition is characterized by

$$\text{Condition B} = \begin{cases} \mathbf{X}_i \in \text{non-ROI}, & \text{and} \\ \mathcal{S}_s(\mathbf{X}_i) > \mathcal{S}_s(\widetilde{\mathbf{X}}_i(\psi|Q_i)). \end{cases}$$

Note that minimizing the saliency distortion in (11) will attempt to preserve the saliency of $\mathbf{X}_i$ after compression. However, as outlined at the beginning of this section, we want to allow saliency change in the desired direction when either Condition A or B holds. Hence, if either Condition A or B holds, the saliency distortion should be removed from the cost function (9), because in these cases the saliency is changing in the desired direction. This is achieved by setting the saliency-related Lagrange multiplier λ_{S_i} to zero:

$$\lambda_{S_i} = \begin{cases} 0, & \text{if Condition A or B holds,} \\ \lambda_S, & \text{otherwise,} \end{cases} \tag{13}$$

where λ_S is a user-defined parameter. Hence, in such cases, the coding mode will be chosen by considering conventional rate and distortion only, while the saliency will be allowed to change in the desired direction – increase in ROI, and decrease in non-ROI. In our experiments, we set $\lambda_S = 1.5$, which worked well for our test sequences. An optimized

未来工作

choice of λ_S, possibly adaptive on a block-by-block basis, is an interesting topic for future research, but is beyond the scope of the present work.

We next discuss the choice of λ_{R_i}. From (9), λ_{R_i} can be obtained by calculating the partial derivative of J_i with respect to R, then setting it to zero, and finally solving for λ_{R_i}. More specifically, we need to have

$$\begin{aligned}&\frac{\partial J_i(\psi|Q_i,\lambda_{S_i},\lambda_{R_i},\mathbf{X}_i)}{\partial R(\psi|Q_i,\mathbf{X}_i)}\\&=\frac{\partial D_{MSE}(\psi|Q_i,\mathbf{X}_i)}{\partial R(\psi|Q_i,\mathbf{X}_i)}+\lambda_{S_i}\frac{\partial D_{sal}(\psi|Q_i,\mathbf{X}_i)}{\partial R(\psi|Q_i,\mathbf{X}_i)}+\lambda_{R_i}\\&=0.\end{aligned}\tag{14}$$

Solving for λ_{R_i} gives

$$\lambda_{R_i}=-\frac{\partial D_{MSE}(\psi|Q_i,\mathbf{X}_i)}{\partial R(\psi|Q_i,\mathbf{X}_i)}-\lambda_{S_i}\frac{\partial D_{sal}(\psi|Q_i,\mathbf{X}_i)}{\partial R(\psi|Q_i,\mathbf{X}_i)}.\tag{15}$$

With Lagrange multipliers set according to (13) and (15), the encoder can choose the optimal coding mode ψ for $\mathbf{X}_i$. We next derive a closed-form expression for λ_{R_i} for the case when the transformed residual of $\mathbf{X}_i$ obeys a Laplacian model.

C. Statistical Modeling of Transformed Residuals

Following [47], we model the marginal density of transformed residuals Y by a zero-mean Laplace probability density function with parameter λ,

$$f_Y(y;\lambda)=\frac{\lambda}{2}e^{-\lambda|y|}.\tag{16}$$

The relationship between λ and standard deviation σ_Y is

$$\lambda=\frac{\sqrt{2}}{\sigma_Y}.\tag{17}$$

To describe the correlation structure of the signal, we adopt a separable autocorrelation function $r_i(m,n)=\sigma_{r_i}^2\rho_i^{|m|}\rho_i^{|n|}$, where m and n are the horizontal and vertical distances between samples, respectively, $\sigma_{r_i}^2$ is the variance of the residual signal of MB $\mathbf{X}_i$ before transformation, and ρ_i is the correlation coefficient of the residual signal of MB $\mathbf{X}_i$, assumed to be equal in horizontal and vertical directions. Using such a model, the variance of the (j,l)-th transform coefficient obtained under coding mode ψ can then be obtained as follows [1], [48], [49]

$$\sigma_{Y_i}^2(j,l)=\sigma_{r_i}^2[\mathbf{A}(\psi)\mathbf{K}_i(\psi)\mathbf{A}(\psi)^T]_{j,j}[\mathbf{A}(\psi)\mathbf{K}_i(\psi)\mathbf{A}(\psi)^T]_{l,l},\tag{18}$$

where $\mathbf{A}(\psi)$ is the $N\times N$ transform matrix for the coding mode ψ and $\mathbf{K}(\psi)$ is the $N\times N$ covariance matrix

$$\mathbf{K}_i(\psi)=\begin{bmatrix}1&\rho_i&\rho_i^2&\cdots&\rho_i^{N-1}\\\rho_i&1&&&\\\rho_i^2&&\ddots&&\vdots\\\vdots&&&&\rho_i\\\rho_i^{N-1}&&\cdots&\rho_i&1\end{bmatrix}.\tag{19}$$

In (18), notation $[.]_{j,j}$ means the (j,j)-th element of the matrix. Hence, according to the adopted model, the (j,l)-th transform coefficient of the residual of $\mathbf{X}_i$ is a Laplacian random variable with parameter

$$\lambda_i^{jl}=\frac{\sqrt{2}}{\sigma_{Y_i}(j,l)}.\tag{20}$$

Note that the correlation coefficient ρ_i and variance $\sigma_{r_i}^2$ are estimated from the residual signal of MB $\mathbf{X}_i$ for each i. Hence, the model is adapted locally to the data.

D. The Rate Model

The rate of MB $\mathbf{X}_i$ is obtained from the entropy of its quantized transformed residual. The entropy of the (j,l)-th coefficient is given by

$$\begin{aligned}h_i(j,l)=&-p_{i_0}(j,l)\log_2 p_{i_0}(j,l)\\&-2\sum_{n=1}^{\infty}p_{i_n}(j,l)\log_2 p_{i_n}(j,l),\end{aligned}\tag{21}$$

where p_{i_0} and p_{i_n} are the probabilities of transformed residuals being quantized to the zeroth and n-th quantization levels, respectively, and can be obtained as

$$p_{i_0}(j,l)=\int_{-(Q_i-\gamma Q_i)}^{(Q_i+\gamma Q_i)}f_{\lambda_i^{jl}}(x)dx,\tag{22}$$

$$p_{i_n}(j,l)=\int_{nQ_i-\gamma Q_i}^{(n+1)Q_i-\gamma Q_i}f_{\lambda_i^{jl}}(x)dx,\tag{23}$$

where Q_i is the quantization step size of $\mathbf{X}_i$, and $F_i=\gamma Q_i$ denotes the rounding offset of the quantizer with $\gamma\in(0,1)$. In H.264/AVC, $\gamma=1/6$ for inter frames and $\gamma=1/3$ for intra frames [15], [16]. The total rate of MB $\mathbf{X}_i$ coded under coding mode ψ with quantization step size Q_i can be estimated from the sum of entropies of individual transform coefficients

$$R(\psi|Q_i,\mathbf{X}_i)=\zeta\sum_{(j,l)}h_i(j,l),\tag{24}$$

where ζ is a factor to compensate for the inaccuracies in the model. For example, the transform coefficients are assumed to be correlated in Section IV-C, which will result in a lower rate than the sum of their individual entropies. Hence, we expect $\zeta<1$. In our experiments, we set $\zeta=0.8$.

In order to simplify subsequent equations, we define the following symbols for commonly used quantities:

$$\begin{aligned}v_i^{jl}&=\lambda_i^{jl}Q_i\\\phi_i^{jl}&=e^{-v_i^{jl}}-1\\\xi_i^{jl}&=e^{\lambda_i^{jl}(F_i-Q_i)}\\\psi_i^{jl}&=e^{v_i^{jl}}\\\kappa_i^{jl}&=(\lambda_i^{jl})^2Q_iF_i\\\theta_i^{jl}&=\lambda_i^{jl}\xi_i^{jl}\\\eta_i^{jl}&=1-e^{\lambda_i^{jl}(F_i-Q_i)}.\end{aligned}\tag{25}$$

Substituting (21)-(23) into (24), using (25), we obtain a closed-form expression for the rate of $\mathbf{X}_i$ in (26).

$$R(\psi|Q_i,\mathbf{X}_i) = -\frac{\zeta}{\ln 2}\sum_{(j,l)}\left(\xi_i^{jl}\left(\ln(-\phi_i^{jl}) - \ln 2 + F_i\lambda_i^{jl}\frac{\nu_i^{jl}}{\phi_i^{jl}}\right) + \eta_i^{jl}\ln(\eta_i^{jl})\right). \quad (26)$$

E. The Distortion Models

The total MSE distortion in $\mathbf{X}_i$ is the sum of quantization distortions contributed by individual transform coefficients

$$D_{MSE}(\psi|Q_i,\mathbf{X}_i) = \sum_{(j,l)}\left(\int_{-(Q_i-\gamma Q_i)}^{(Q_i+\gamma Q_i)} x^2 f_{\lambda_i^{jl}}(x)dx + 2\sum_{n=1}^{\infty}\int_{nQ_i-\gamma Q_i}^{(n+1)Q_i-\gamma Q_i}(x-nQ_i)^2 f_{\lambda_i^{jl}}(x)dx\right). \quad (27)$$

After some algebraic manipulation, $D_{MSE}(\psi|Q_i,\mathbf{X}_i)$ can be expressed in the closed form as in (29), shown at the bottom of the page.

Based on (11), the saliency distortion of a MB is proportional to the spatial saliency distortion of the MB weighted by the motion saliency of the MB. As described in Section III, our approximation to the spatial saliency of a MB is the energy of the Wiener-filtered DCT of the MB. In order to estimate the spatial saliency distortion of a block due to quantization, we model the quantization process by an equivalent quantization noise [50], and consider the Wiener-weighted energy of the quantization noise in the DCT domain as our spatial saliency distortion. More specifically, we consider the following closed-form expression as our proposed model for $D_{sal}(\psi|Q_i,\mathbf{X}_i)$.

$$D_{sal}(\psi|Q_i,\mathbf{X}_i) = \mu_i\sum_{(j,l)}\mathbf{H}(j,l)\chi_i^{jl}, \quad (28)$$

where χ_i^{jl} is as defined in (29), shown at the bottom of the page.

F. A Closed-Form Expression for λ_{R_i}

From the expressions for $R(\psi|Q_i,\mathbf{X}_i)$, $D_{MSE}(\psi|Q_i,\mathbf{X}_i)$ and $D_{sal}(\psi|Q_i,\mathbf{X}_i)$, we can now obtain the expression for λ_{R_i}. To do this, using the chain rule, we express the ratios in (15) in terms of partial derivatives with respect to Q_i,

$$\lambda_{R_i} = -\frac{\frac{\partial}{\partial Q_i}(D_{MSE}(\psi|Q_i,\mathbf{X}_i) + \lambda_{S_i}D_{sal}(\psi|Q_i,\mathbf{X}_i))}{\frac{\partial R(\psi|Q_i,\mathbf{X}_i)}{\partial Q_i}}, \quad (30)$$

where the numerator is given in (31), and the denominator is given in (32), both shown at the bottom of the page.

Note that several quantities, such as ν_i^{jl} and λ_i^{jl}, appear in both (31) and (32), which means that computational effort can be reduced by computing these quantities only once. In our implementation, we first compute (31), and values of the quantities that are shared with (32) are reused.

It is worth pointing out that λ_{R_i} in (30) depends on the content of each MB through the variance and correlation of the residual of MB $\mathbf{X}_i$, as well as the motion and spatial saliency of $\mathbf{X}_i$, based on (11). Hence, using (30), we can adjust the Lagrange multiplier in a content-adaptive manner on a MB-by-MB basis.

V. Experimental Results

In this section, we first evaluate the performance of the proposed GMC saliency estimation method in Section V-A. We then evaluate the performance of the proposed saliency-aware video compression method in Section V-B. The computational complexity of the proposed video compression method is discussed in Section V-D.

A. Evaluating the proposed GMC saliency estimation method

In order to evaluate the performance of the GMC saliency estimation method proposed in Section III-B, we compared it against the IKN model on the eye-tracking dataset from [51]. This dataset contains eye-tracking data for two viewings of 12 standard test video sequences. To generate the results for the IKN model, the original implementation of the IKN model

$$D_{MSE}(\psi|Q_i,\mathbf{X}_i) = \sum_{(j,l)}\chi_i^{jl} = \sum_{(j,l)}\frac{e^{\lambda_i^{jl}F_i}(2\nu_i^{jl} + (\nu_i^{jl})^2 - 2\kappa_i^{jl}) + 2 - 2\psi_i^{jl}}{-\phi_i^{jl}(\lambda_i^{jl})^2}. \quad (29)$$

$$\frac{\partial}{\partial Q_i}(D_{MSE}(\psi|Q_i,\mathbf{X}_i) + \lambda_{S_i}D_{sal}(\psi|Q_i,\mathbf{X}_i)) = \sum_{(j,l)}(1+\mu_i\lambda_s\mathbf{H}(j,l))\left(\frac{2\lambda_i^{jl}\nu_i^{jl} - e^{F_i\lambda_i^{jl}}\left(2\lambda_i^{jl} - 2F_i(\lambda_i^{jl})^2 + 2(\lambda_i^{jl})^2Q_i\right)}{(\lambda_i^{jl})^2(\psi_i^{jl}-1)} + \frac{\psi_i^{jl}\left(e^{F_i\lambda_i^{jl}}\left((\nu_i^{jl})^2 - 2F_i(\lambda_i^{jl})^2Q_i + 2\lambda_i^{jl}Q_i\right) - 2\nu_i^{jl} + 2\right)}{\lambda_i^{jl}(\psi_i^{jl}-1)^2}\right) \quad (31)$$

$$\frac{\partial R(\psi|Q_i,\mathbf{X}_i)}{\partial Q_i} = \frac{-\zeta}{\ln 2}\sum_{(j,l)}\theta_i^{jl}\left(1 + \left(\frac{\lambda_i^{jl}}{\phi_i^{jl}} - \frac{\lambda_i^{jl}}{\psi_i^{jl}\phi_i^{jl}} + \frac{(\lambda_i^{jl})^2Q_i}{\psi_i^{jl}(\phi_i^{jl})^2}\right) + \ln(-\phi_i^{jl}) - \left(\ln(-\phi_i^{jl}) - \ln 2 + F_i\lambda_i^{jl} + \frac{\nu_i^{jl}}{\phi_i^{jl}}\right)\right) \quad (32)$$

[52] was utilized. Note that as introduced in [12], in the original implementation of the IKN model for video, two main normalization operators are available for combining the conspicuity and feature maps: Maxnorm and Fancyone. Maxnorm yields smoother, more continuous saliency maps, while Fancyone yields increasingly sparser saliency maps, with only a few sharp peaks [12]. Since the saliency maps produced by Maxnorm are smoother than those of Fancyone, the Maxnorm operator is thought to be better suited to video compression [12]. For the sake of simplicity in the rest of our analysis, we call the IKN model with the Maxnorm operator "IKN-MA," and the IKN model with Fancyone "IKN-FA."

To find appropriate values for α and β in (6), we varied α in the range $[0.5, 1]$ with a step size of 0.1, and β in the range $[0.1, 1]$ with a step size of 0.1. For each combination of α and β, we measured the average eye-tracking score [51] of the proposed saliency estimation method over all videos in the data set in [51]. The overall average eye-tracking score was the highest with $\alpha = 0.9$ and $\beta = 1$, although the performance was not too sensitive to these parameters. Hence, we used $\alpha = 0.9$ and $\beta = 1$ in our experiments, but other nearby values would have been appropriate as well. An alternative way to find possibly better values for α and β would be to use machine learning techniques with independent datasets for training and testing similar to [53] and [54].

Table I compares the proposed saliency estimation method with IKN-MA based on the average accuracy score for predicting gaze locations as defined in [51] for both viewings of each of the 12 sequences used in [51]. As seen from these results, in all cases the average accuracy score of our proposed method is higher than that of IKN-MA. To examine whether the difference in the average scores between IKN-MA and our method is statistically significant, we performed a paired t-test [55] on the frame-by-frame scores for each sequence and each viewing. The null hypothesis was that the scores of both models come from the distributions with the same mean. Based on these results, we observe that the obtained p-values were less than 2×10^{-3}, except for *Crew* for both viewings in which the null hypothesis was not rejected due to a p-value larger than 0.05. This means that in all cases except for *Crew*, the average accuracy score of the proposed saliency estimation method was higher than IKN-MA, and these results were statistically significant due to a very small p-value. However, there was a statistical tie between the two models on *Crew*. Hence, based on this data, one could conclude that the proposed saliency estimation method is more accurate than IKN-MA.

In order to compare the accuracy of our proposed saliency estimation method with IKN-FA, we first applied the Fancyone normalization operator on all saliency maps produced by our proposed saliency estimation method. We then performed the same analysis as in Table I based on the obtained saliency maps. The results are reported in Table II. In this table, p-values larger than 0.05 have been indicated in bold typeset. According to the results in Table II, we observe that the performance of the proposed saliency estimation method is statistically the same as IKN-FA on *Bus* (for both viewings), *Crew* (for both viewings), *Flower Garden* (for the second viewing), *Harbor* (for both viewings), *Stefan* (for both viewings), and *Tempete* (for the second viewing). In these cases, the null hypothesis cannot be rejected because the corresponding p-values are larger than 0.05. In all other cases, the proposed saliency estimation method outperforms IKN-FA except for the first viewing of *Flower Garden*, and both viewings of *Hall Monitor*, *Mobile Calendar*, and *Mother & Daughter*. Hence, one could argue that the proposed saliency estimation method results in comparable results with IKN-FA.

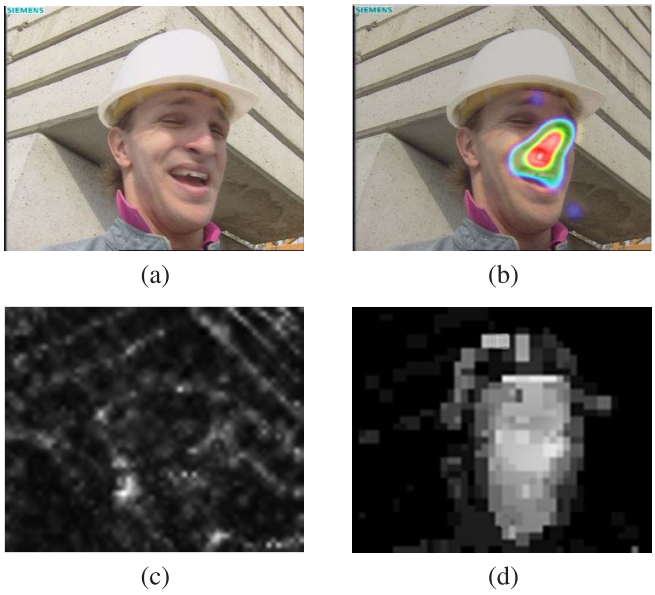


Fig. 1. A frame from *Foreman*: (a) original frame (b) heat map of the actual gaze locations (c) the saliency map generated by the IKN model (d) the saliency map generated by the proposed method.

Fig. 1 shows a frame of *Foreman* along with the heat map generated from the actual gaze locations as given by the eye-tracking database in [51] as well as the saliency maps generated by the IKN model and the proposed saliency estimation method for video. As seen from this example, the proposed method is able to find the main salient foreground region in the scene.

B. Quantitative evaluation of the proposed video compression method

In order to measure the subjective quality of the proposed saliency-aware video compression method, we used the eye-tracking weighted mean square error (EWMSE) metric proposed in [6]. The EWMSE value of a encoded video frame can be computed as follows [6]

$$EWMSE = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H}\left(w_{x,y}\cdot(F'_{x,y}-F_{x,y})^2\right)}{WH\sum_{x=1}^{W}\sum_{y=1}^{H}w_{x,y}}, \quad (33)$$

where $F'_{x,y}$ and $F_{x,y}$ respectively denote the pixel at location (x, y) in the encoded frame $\mathbf{F}'$ and the original frame $\mathbf{F}$, W and H are the width and height of $\mathbf{F}$ in pixels, and $w_{x,y}$ is the weight for distortion at pixel location (x, y), and it is obtained

TABLE I

COMPARING THE PROPOSED SALIENCY ESTIMATION METHOD WITH THE IKN MODEL BASED ON THE EYE-TRACKING DATABASE IN [51] WHEN THE MAXNORM OPERATOR IS EMPLOYED IN THE IKN MODEL.

Video	*First Viewing*				*Second Viewing*			
	IKN-MA	Proposed	p-value	Difference	IKN-MA	Proposed	p-value	Difference
Bus	19.18	22.30	0.000006	+3.12	19.35	21.81	0.002081	+2.46
City	15.86	33.05	0.000000	+17.19	16.04	33.77	0.000000	+17.73
Crew	15.66	15.87	**0.323377**	+0.21	15.93	15.98	**0.775784**	+0.05
Foreman	20.05	22.71	0.000405	+2.66	19.99	23.67	0.000003	+3.68
Flower Garden	20.69	22.07	0.000000	+1.38	21.48	23.03	0.000000	+1.55
Hall Monitor	30.47	34.12	0.000245	+3.65	32.18	38.92	0.000000	+6.74
Harbor	18.83	21.27	0.001148	+2.44	19.37	22.06	0.000065	+2.69
Mobile Calendar	19.42	26.09	0.000000	+6.67	19.43	28.06	0.000000	+8.63
Mother & Daughter	15.51	16.78	0.000000	+1.27	16.03	17.06	0.000000	+1.03
Soccer	18.81	21.62	0.000000	+2.81	20.20	24.73	0.000000	+4.53
Stefan	21.07	23.97	0.007764	+2.90	18.67	23.05	0.000189	+4.38
Tempete	17.96	21.42	0.000000	+3.46	17.36	20.54	0.000000	+3.18

TABLE II

COMPARING THE PROPOSED SALIENCY ESTIMATION METHOD WITH THE IKN MODEL BASED ON THE EYE-TRACKING DATABASE IN [51] WHEN FANCYONE OPERATOR IS EMPLOYED

Video	*First Viewing*				*Second Viewing*			
	IKN-FA	Proposed	p-value	Difference	IKN-FA	Proposed	p-value	Difference
Bus	23.50	21.33	**0.161670**	−2.17	20.83	19.23	**0.365476**	−1.6
City	9.12	68.04	0.000000	+58.92	10.67	62.47	0.000000	+51.08
Crew	19.60	20.17	**0.598369**	+0.57	18.95	18.83	**0.891117**	−0.12
Foreman	28.99	38.40	0.000000	+9.41	30.01	41.62	0.000000	+11.61
Flower Garden	51.31	44.71	0.000000	−6.60	48.93	47.60	**0.371274**	−1.33
Hall Monitor	81.62	47.31	0.000000	−34.31	83.71	55.79	0.000000	−27.92
Harbor	31.12	33.14	**0.279229**	+2.02	36.81	34.29	**0.093973**	−2.52
Mobile Calendar	44.74	31.55	0.000000	−13.19	40.21	34.46	0.000708	−5.75
Mother & Daughter	32.63	18.60	0.000000	−14.03	35.13	17.84	0.000000	−17.29
Soccer	31.19	39.68	0.000000	+8.49	29.12	40.15	0.000000	+11.03
Stefan	67.42	73.92	**0.231836**	+6.50	66.91	73.50	**0.242396**	+6.59
Tempete	34.33	37.18	0.040270	+2.85	34.05	31.70	**0.086432**	−2.735

using the following 2-D Gaussian function

$$w_{x,y} = \frac{1}{2\pi\sigma_x\sigma_y G}\sum_{g=1}^{G} e^{-\left(\frac{(x-x_{p_g})^2}{2\sigma_x^2}+\frac{(y-y_{p_g})^2}{2\sigma_y^2}\right)}, \qquad (34)$$

where (x_{p_g}, y_{p_g}) is the eye fixation position of the g-th subject, which is given by our eye-tracking database in [51], $G = 15$ is the total number of subjects in the eye-tracking database, and σ_x and σ_y are two parameters that specify the extent or width of the Gaussian function, and they depend on the viewing distance and view angle. The values of σ_x and σ_y can be taken based on the fovea size, which is about $2-5^\circ$ of visual angle [6], [51]. Here, similar to [6] and [51] we use $\sigma_x = \sigma_y = 64$ pixels, which is equivalent to 2° of visual angle. Using the EWPSNR metric given by (33), we can now define an equivalent eye-tracking-weighted PSNR as follows

$$EWPSNR = 10\log_{10}\left(\frac{255^2}{EWMSE}\right). \qquad (35)$$

In this set of experiments, we considered the mean EWPSNR of a video as a measure of the subjective quality of the video. Hence, the higher the EWPSNR, the better the subjective quality of the encoded video.

In order to evaluate the performance of the proposed saliency-aware video compression method, we compared it based on the mean EWPSNR at several bitrates with the conventional RDO method implemented in the H.264/AVC reference software JM 16.1 [15], as well as two recent ROI-coding methods: the FJND method proposed in [23] and the visual attention guided bit allocation (VAGBA) method proposed in [6]. For this purpose, we used the 12 standard CIF sequences from [51] whose EWPSNR can be computed using the eye-tracking data provided in [51]. Better average EWPSNR values are expected if, on average, the predictions of the saliency model are closer to the actual human fixation points. To compute the EWPSNR values, we used the first viewing eye-tracking data provided in [51], and only the luma channel of the videos was used.

All videos were encoded by each of the aforementioned methods at different bit rates with a GOP structure of IPPP. To encode a video at different bit rates, we varied the frame-level QP (QP_f) of the video between 25 to 40, and at each value, we computed the average EWPSNR and PSNR of the encoded video. This results in a RD curve for each video encoded using one of the above-mentioned coding methods. In Fig. 2 and 3, we plotted the RD curves for *Foreman* and *Tempete* in which the distortion is quantified by means of EWPSNR.

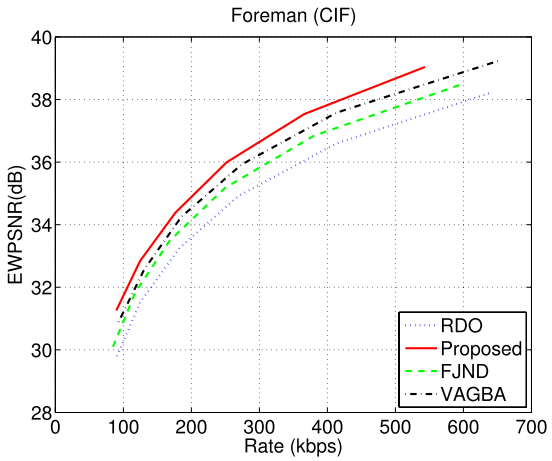


Fig. 2. A plot of EWPSNR versus rate for *Foreman*.

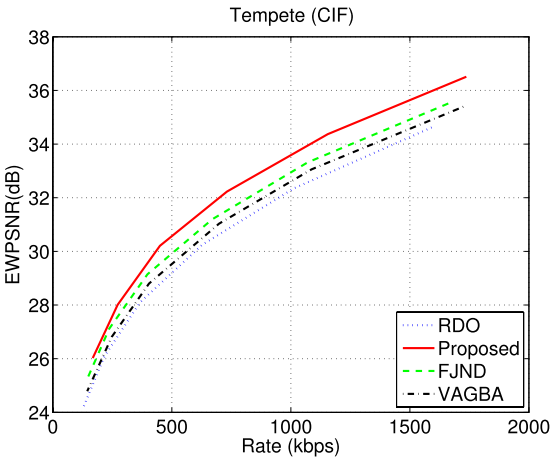


Fig. 3. A plot of EWPSNR versus rate for *Tempete*.

We utilized the well-known BD-PSNR [56] metric to measure the average difference between two RD curves in terms of PSNR difference. Similarly, we computed the BD-EWPSNR values of the encoded videos. In order to be able to compare the RD performance of various methods, we considered the conventional RDO method as our baseline and computed the BD-PSNR and BD-EWPSNR of the other three methods with respect to the conventional RDO method.

We first compare the various methods in terms of their bit allocation strategy. To do this, we remove the influence of saliency estimation and its accuracy by using saliency maps produced by eye-tracking heatmaps from [51]. This way, FJND, VAGBA, and the proposed method use the same saliency maps, which in turn precisely match the eye-tracking data, leaving bit allocation as the main difference among the methods. Table III shows BD-EWPSNR and BD-PSNR results with conventional RDO method taken as the benchmark. As seen from the results, the proposed method is able to provide an average EWPSNR gain of 2.05 dB with respect to RDO, 1.00 dB (= 2.05 dB −1.05 dB) with respect to VAGBA, and 0.67 dB (= 2.05 dB −1.38 dB) with respect to FJND. In terms of conventional PSNR, the average gain of the proposed method is 0.25 dB (= −0.01 dB +0.26 dB) with respect to FJND, and 0.14 dB (= −0.01 dB +0.15 dB) with respect to VAGBA, while the average loss against RDO is minimal (0.01 dB). These results indicate that the bit allocation strategy of the proposed method is more efficient than that of FJND and VAGBA.

Next, we compare the combination of the proposed methods (that is, the proposed video coding method coupled with the proposed saliency estimation) against the state of the art. As the state of the art, we take FJND and VAGBA coupled with the IKN saliency model, which was recently shown as the most accurate in terms of gaze prediction on video among the nine tested methods in [33] on the eye-tracking dataset from [51]. We also included our previous saliency-preserving (SP) video coding method from [10] in this comparison. The methods are compared in terms of BD-EWPNSR, BD-PSNR, BD-SSIM [24], and BD-VQM [57], [58] in Table IV, with RDO taken as the benchmark. Note that the lower the VQM value, the higher the visual quality measured by VQM. As seen from the table, on average, the proposed methods increase the BD-EWPSNR by 1.45 dB with respect to conventional RDO, at the same BD-PSNR. Moreover, the proposed methods improve the video quality in terms of all metrics (EWPSNR, PSNR, SSIM, VQM) compared to FJND, VAGBA, and SP.

It is interesting to note that RDO performs slightly better, on average, than any of the perceptually-motivated ROI-based video coding methods in terms of SSIM and VQM. This is likely due to the fact that both SSIM and VQM ignore visual attention (i.e., saliency), while SSIM in addition ignores temporal quality.

The above-mentioned methods are all ROI-based video coding methods that use saliency information for video compression. In order to see the benefit of using visual saliency for video compression, we measured the performance of a SSIM-based video coding method [59] in terms of BD-EWPNSR, BD-PSNR, BD-SSIM, and BD-VQM. The method in [59] does not use visual saliency. The results are shown in Table V. Comparing the results in Table V with the results in Table IV, we observe that the performance of the proposed methods is better than that of [59] in terms of BD-EWPSNR and BD-PSNR, but lower in terms of BD-SSIM and BD-VQM. This is again likely due to the fact that both SSIM and VQM do not consider saliency. However, as will be seen in Section V-C, the subjective quality of videos encoded by the proposed methods, as judged by human viewers, is better than that of [59].

C. Subjective Evaluation

Finally, we performed a subjective evaluation of the perceptual quality of sequences encoded using the proposed saliency-aware compression method versus sequences encoded using FJND [23]. We chose FJND as the competing method here because it performed slightly better than VAGBA in the tests described above. We utilized a Two Alternative Forced Choice (2AFC) method [60] to compare subjective video quality. In 2AFC, the participant is asked to make a choice between two alternatives, in this case, the video encoded using the proposed method vs. video encoded using FJND. This way

TABLE III
COMPARING THE PROPOSED VIDEO COMPRESSION METHOD WITH THE FJND METHOD [23] AND THE VAGBA METHOD [6] BASED ON THE AVERAGE BD-EWPSNR AND BD-PSNR VALUES WITH RESPECT TO THE CONVENTIONAL RDO METHOD WHEN THE EYE-TRACKING HEATMAPS ARE USED AS THE SALIENCY MAPS

Video	*FJND* [23]		*VAGBA* [6]		*Proposed*	
	BD-EWPSNR	BD-PSNR	BD-EWPSNR	BD-PSNR	BD-EWPSNR	BD-PSNR
Bus	+1.58	−0.21	+1.16	−0.15	+2.16	+0.08
City	+1.40	−0.27	+1.01	−0.15	+1.98	−0.15
Crew	+1.53	−0.10	+1.02	−0.07	+1.90	+0.20
Foreman	+1.52	−0.31	+0.89	−0.19	+1.98	−0.17
Flower Garden	+1.77	−0.25	+1.26	−0.11	+2.31	+0.02
Hall Monitor	+1.33	−0.54	+0.94	−0.29	+1.79	−0.33
Harbor	+1.02	−0.19	+1.07	−0.14	+1.95	−0.16
Mobile Calendar	+1.00	−0.18	+1.15	−0.13	+2.39	+0.31
Mother & Daughter	+1.28	−0.37	+0.91	−0.18	+2.03	−0.31
Soccer	+1.01	−0.13	+0.93	−0.14	+1.88	−0.05
Stefan	+1.80	−0.39	+1.13	−0.16	+2.04	−0.04
Tempete	+1.67	−0.15	+1.15	−0.09	+2.21	+0.43
Average	+1.38	−0.26	+1.05	−0.15	+2.05	−0.01

of comparing quality is less susceptible to measurement noise than quality ratings based on scale, such as Mean Opinion Score (MOS) and Double Stimulus Continuous Quality Scale (DSCQS) [61], because participant's task is much simpler than mapping quality to a number on the scale.

All 12 CIF sequences from the dataset [51] were used in the experiment. All sequences were encoded with a GOP structure of IPPP using the two compression methods. The average PSNR of the encoded videos was around 31 dB. In each trial, participants were shown two videos, side by side, at the same vertical position separated by 1 cm horizontally on a mid-gray background. Each video pair was shown for 10 seconds. After this presentation, a mid-gray blank screen was shown for 5 seconds. During this period, participants were asked to indicate on an answer sheet, which of the two videos looks better (Left or Right). They were asked to answer either Left or Right for each video pair, regardless of how certain they were of their response. Participants did not know which video was produced by the proposed method and which one was produced by FJND. Randomly chosen half of the trials had the video produced by the proposed method on the left side of the screen and the other half on the right side, in order to counteract side bias in the responses. This gave a total of $12 \cdot 2 = 24$ trials.

The experiment was run in a quiet room with 15 participants (14 male, 1 female, aged between 18 and 30). All participants had normal or corrected to normal vision. A 22-inch Dell monitor with brightness 300 cd/m^2 and resolution 1680 × 1050 pixels was used in our experiments. The brightness and contrast of the monitor were set to 75%. The actual height of the displayed videos on the screen was 185 millimeters. The illumination in the room was in the range 280–300 Lux. The distance between the monitor and the subjects was fixed at 80 cm. Each participant was familiarized with the task before the start of the experiment via a short printed instruction sheet. The total length of the experiment for each participant was approximately 6 minutes.

The results for the comparison are shown in Table VI. In Table VI we show the number of responses that showed preference for the FJND method vs. the proposed method.

To test for statistical significance, we used a two-sided χ^2-test [62], with the null hypothesis that there is no preference for either method, i.e., that the votes for each method come from distributions with the same mean. Under this hypothesis, the expected number of votes in each trial is 15 for each method, because each video pair was shown twice to each of the 15 participants. The p-value [62] of the test is indicated in the table. As a rule of thumb, the null hypothesis is rejected when $p < 0.05$. When this happens in Table VI, it means that the two methods under the comparison cannot be considered to have the same subjective quality, since one of them has obtained a statistically significantly higher number of votes, and therefore seems to have better quality.

In 8 out of the 12 cases in Table VI we have $p < 0.05$, which indicates that subjects showed a statistically significant preference for the proposed method vs. FJND. In only 4 cases (*Bus*, *Flower Garden*, *Harbor*, and *Mobile Calendar*) the p-value is larger than 0.05, which means that neither method achieved a statistically significant advantage. Looking across all trials (i.e., summing up all the votes for the two methods), the results show that participants have preferred the proposed method much more than FJND (268 vs. 92 votes) with overall $p = 0.0001$, which is a very statistically significant result. This confirms that the proposed method is able to provide higher perceptual video quality compared to FJND.

We also performed a subjective evaluation of the quality of sequences encoded by the proposed saliency-aware video compression method versus sequences encoded by the SSIM-based video coding method from [59] using the experimental procedure described above. The results are shown in Table VII. As seen from these results, the proposed

TABLE IV

COMPARING VARIOUS METHODS WITH CONVENTIONAL RDO BASED ON THE AVERAGE BD-EWPSNR, AVERAGE BD-PSNR, AVERAGE BD-SSIM, AND AVERAGE BD-VQM VALUES

FJND [23]				
Video	BD-EWPSNR	BD-PSNR	BD-SSIM	BD-VQM
Bus	+0.28	−0.18	−0.007362	+0.011053
City	+0.07	−0.09	−0.003250	+0.004384
Crew	+0.27	−0.08	−0.002006	+0.009742
Foreman	+0.19	−0.16	−0.002003	+0.006101
Flower Garden	+0.60	−0.12	−0.002202	+0.007403
Hall Monitor	+0.67	−0.12	−0.002378	+0.014506
Harbor	+0.13	−0.15	−0.004979	+0.017908
Mobile Calendar	+0.21	−0.12	−0.003537	+0.009658
Mother & Daughter	+0.46	−0.29	−0.004065	+0.016102
Soccer	+0.23	−0.17	+0.024222	−0.008660
Stefan	+0.65	−0.12	−0.000593	+0.000102
Tempete	+0.98	−0.09	−0.003144	+0.009854
Average	+0.40	−0.15	−0.003484	+0.010925
VAGBA [6]				
Video	BD-EWPSNR	BD-PSNR	BD-SSIM	BD-VQM
Bus	+0.37	−0.14	−0.005982	+0.009714
City	+0.06	−0.15	−0.005329	+0.012298
Crew	+0.49	−0.13	−0.003320	+0.017451
Foreman	+0.48	−0.24	−0.003009	+0.011736
Flower Garden	+0.81	−0.13	−0.002743	+0.013063
Hall Monitor	+0.81	−0.13	−0.003006	+0.031234
Harbor	+0.28	−0.16	−0.004281	+0.011610
Mobile Calendar	+0.32	−0.13	−0.003245	+0.004019
Mother & Daughter	+0.32	−0.23	−0.002226	+0.006420
Soccer	+0.54	−0.23	−0.008660	+0.021080
Stefan	+0.67	−0.13	−0.000863	+0.000119
Tempete	+0.71	−0.13	−0.002862	+0.009357
Average	+0.49	−0.17	−0.003794	+0.012347
SP [10]				
Video	BD-EWPSNR	BD-PSNR	BD-SSIM	BD-VQM
Bus	+0.41	−0.12	−0.002982	+0.001561
City	+0.71	−0.32	−0.015125	+0.029112
Crew	+0.45	−0.12	−0.001002	+0.006526
Foreman	+0.81	−0.22	−0.001305	+0.007993
Flower Garden	+0.89	−0.12	+0.002356	−0.001473
Hall Monitor	+0.78	−0.16	−0.002431	+0.018221
Harbor	+0.53	−0.23	−0.005926	+0.0285378
Mobile Calendar	+0.77	−0.27	+0.004436	−0.004334
Mother & Daughter	+0.64	−0.31	−0.003441	+0.015512
Soccer	+0.80	−0.42	−0.013345	+0.035527
Stefan	+0.77	−0.12	+0.001551	−0.000221
Tempete	+0.72	−0.11	+0.003342	−0.001662
Average	+0.69	−0.21	−0.002802	+0.011301
Proposed				
Video	BD-EWPSNR	BD-PSNR	BD-SSIM	BD-VQM
Bus	+0.93	+0.02	−0.002766	+0.001434
City	+1.55	−0.27	−0.014254	+0.028715
Crew	+0.94	+0.11	+0.001360	+0.005731
Foreman	+1.62	−0.11	−0.001289	+0.007552
Flower Garden	+1.73	+0.14	+0.003143	−0.001978
Hall Monitor	+1.65	−0.09	−0.002385	+0.017588
Harbor	+0.98	−0.12	−0.005621	+0.022322
Mobile Calendar	+1.50	+0.38	+0.006875	−0.006075
Mother & Daughter	+1.54	−0.26	−0.004703	+0.017603
Soccer	+1.30	−0.31	−0.014136	+0.036290
Stefan	+1.67	+0.07	+0.001622	−0.000051
Tempete	+1.95	+0.45	+0.007581	−0.011511
Average	+1.45	+0.00	−0.002048	+0.009802

method offers higher subjective quality compared to the method from [59] on most sequences. The exceptions were *Bus* and *Flower Garden*, where there was a statistical tie.

D. Computational Complexity

The proposed saliency-aware video compression method was implemented in the H.264/AVC reference software JM 16.1 [15]. The average encoding time per CIF resolution

TABLE V
COMPARING [59] WITH CONVENTIONAL RDO BASED ON THE AVERAGE BD-EWPSNR, AVERAGE BD-PSNR, AVERAGE BD-SSIM, AND AVERAGE BD-VQM VALUES

[59]				
Video	BD-EWPSNR	BD-PSNR	BD-SSIM	BD-VQM
Bus	−1.10	−0.59	+0.015393	−0.015395
City	+0.10	+0.12	+0.005349	−0.015347
Crew	−0.90	−1.01	−0.003906	+0.014691
Foreman	+0.12	−0.71	−0.002572	−0.005259
Flower Garden	−1.87	−1.65	−0.007392	−0.006491
Hall Monitor	−1.86	−0.70	+0.004971	−0.033708
Harbor	−0.16	−0.13	+0.001502	−0.001101
Mobile Calendar	+0.79	+0.77	+0.014941	−0.038764
Mother & Daughter	−1.10	−0.33	+0.002723	−0.001972
Soccer	−2.11	−1.37	+0.012689	+0.008201
Stefan	+0.18	−0.52	+0.002156	−0.000117
Tempete	+0.10	+0.50	+0.015569	−0.046985
Average	−0.65	−0.46	+0.005100	−0.011900

TABLE VI
SUBJECTIVE COMPARISON OF THE PROPOSED VIDEO COMPRESSION METHOD AGAINST FJND

Sequence	FJND	Proposed	p-value
Bus	12	18	**0.2733**
City	4	26	0.0001
Crew	7	23	0.0035
Foreman	8	22	0.0106
Flower Garden	10	20	**0.0679**
Hall Monitor	9	21	0.0285
Harbor	11	19	**0.1441**
Mobile Calendar	10	20	**0.0679**
Mother & Daughter	4	26	0.0001
Soccer	5	25	0.0003
Stefan	4	26	0.0001
Tempete	8	22	0.0106
Total	92	268	0.0001

TABLE VII
SUBJECTIVE COMPARISON OF THE PROPOSED VIDEO COMPRESSION METHOD AGAINST [59]

Sequence	[59]	Proposed	p-value
Bus	14	16	**0.7150**
City	4	26	0.0001
Crew	6	24	0.0010
Foreman	8	22	0.0106
Flower Garden	12	18	**0.2733**
Hall Monitor	8	22	0.0106
Harbor	5	25	0.0003
Mobile Calendar	9	21	0.0285
Mother & Daughter	6	24	0.0010
Soccer	8	22	0.0106
Stefan	4	26	0.0001
Tempete	4	26	0.0001
Total	88	272	0.0001

frame was about 1.8 seconds for the proposed method on an Intel Core 2 Duo CPU @3.33 GHz with 8 GB RAM. As a comparison, the average encoding time per CIF frame of our previous method from [10] was about 13.4 seconds, while the average encoding time per CIF frame for conventional RDO was 0.6 second . Hence, although the proposed method is about 3 times slower than conventional RDO on our test system, it is about 7 times faster than our previous method from [10].

VI. CONCLUSION

In this paper, we presented a saliency-aware video compression method in the context of ROI-based video coding. The proposed method attempts to reduce attention-grabbing coding artifacts, and further allows the saliency of the encoded video to change in a controlled manner – increase in ROI and decrease in non-ROI. This is achieved by selectively adding a saliency distortion term to the distortion metric used in H.264/AVC rate distortion optimization (RDO). The saliency is estimated through a novel computationally efficient method in the DCT domain, which is inspired by the well-known Itti-Koch-Niebur (IKN) saliency model, but incorporates improved temporal saliency estimation that considers the effect of camera motion. The results indicate that the proposed method is able to improve the visual quality of encoded video compared to conventional RDO video coding, as well as several state-of-the-art perceptually-motivated video coding methods. The framework could be further enhanced by optimizing the saliency-related Lagrange parameter, possibly on a block-by-block basis. Another enhancement could be the inclusion of a more perceptually-friendly distortion metric (e.g., a JND- or SSIM-based distortion metric) instead of the conventional MSE in the RDO cost function.

目的 方法 度量

效果

未来工作

REFERENCES

[1] M. A. Robertson and R. L. Stevenson, "DCT quantization noise in compressed images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 27–38, Jan. 2005.
[2] A. Leontaris, P. C. Cosman, and A. R. Reibman, "Quality evaluation of motion-compensated edge artifacts in compressed video," *IEEE Trans. Image Process.*, vol. 16, no. 11, pp. 943–956, Apr. 2007.
[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[4] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[5] Z. Chen, N. K. Ngan, and W. Lin, "Perceptual video coding: Challenges and approaches," in *Proc. IEEE ICME*, Jul. 2010, pp. 784–789.

[6] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image and Vision Computing*, vol. 29, no. 1, pp. 1–14, 2011.

[7] Y. Liu, Z. Li, and Y. C. Soh, "Region-of-interest based resource allocation for conversational video communications of H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 134–139, Jan. 2008.

[8] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA, USA: MIT Press, 1993, pp. 179–206.

[9] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 011006.1–011006.21, 2010.

[10] H. Hadizadeh and I. V. Bajić, "Saliency-preserving video compression," in *Proc. IEEE ICME*, Jul. 2011, pp. 1–6.

[11] I. E. Richardson, *The H.264 Advanced Video Compression Standard*. New York, NY, USA: Wiley, 2010.

[12] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[13] T. Wiegand and B. Girod, "Lagrange multiplier selection in hybrid video coder control," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3. Oct. 2001, pp. 542–545.

[14] G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 74–90, Nov. 1998.

[15] *The H.264/AVC JM Reference Software* ver. 16.1 [Online] Available: http://iphome.hhi.de/suehring/tml/

[16] X. Li, N. Oertel, A. Hutter, and A. Kaup, "Laplace distribution based Lagrangian rate distortion optimization for hybrid video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 193–205, Feb. 2009.

[17] X. Li, N. Oertel, A. Hutter, and A. Kaup, "Advanced lagrange multiplier selection for hybrid video coding," in *Proc. IEEE ICME*, Jul. 2007, pp. 364–367.

[18] L. Chen and I. Garbacea, "~~Adaptive λ estimation in Lagrangian rate-distortion optimization for video coding~~," *Proc. SPIE*, vol. 6077, pp. 60772B-1–60772B-8, Jan. 2006.

[19] J. Zhang, X. Yi, N. Ling, and W. Shang, "Context adaptive Lagrange multiplier (CALM) for rate-distortion optimal motion estimation in video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 820–828, Jun. 2010.

[20] M. Jiang and N. Ling, "On Lagrange multiplier and quantizer adjustment for H.264 frame-layer video rate control," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 5, pp. 663–669, May 2006.

[21] M. Wang and B. Yan, "Lagrangian multiplier based joint three-layer rate control for H.264/AVC," *IEEE Signal Process. Lett.*, vol. 16, no. 8, pp. 679–682, Aug. 2009.

[22] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, "Motion-compensated residue pre-processing in video coding based on just-noticeable distortion profile," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 6, pp. 742–752, Jun. 2005.

[23] Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 806–819, Jun. 2010.

[24] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[25] C. Yang, R. Leung, L. Po, and Z. Mai, "An SSIM-optimal H.264/AVC inter frame encoder," in *Proc. IEEE Int. Conf. Intell. Comput. Intell. Syst.*, vol. 4. Nov. 2009, pp. 291–295.

[26] C. Yang, H. Wang, and L. Po, "Improved inter prediction based on structural similarity in H.264,," in *Proc. IEEE Int. Conf. Signal Process. Commun.*, vol. 2. Nov. 2007, pp. 340–343.

[27] Y. H. Huang, T. S. Ou, P. Y. Su, and H. Chen, "Perceptual rate-distortion optimization using structural similarity index as quality metric," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1614–1624, Nov. 2010.

[28] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-motivated rate-distortion optimization for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 4, pp. 516–529, Apr. 2012.

[29] X. Wang, L. Su, Q. Huang, and C. Liu, "Visual perception based Lagrangian rate distortion optimization for video coding," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 1653–1656.

[30] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Perceptual video coding based on SSIM-inspired divisive normalization," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1418–1429, Apr. 2013.

[31] C. Yeo, H. L. Tan, and Y. H. Tan, "On rate distortion optimization using SSIM," in *Proc. IEEE ICASSP*, Mar. 2012, pp. 833–836.

[32] X. Yang, W. Lin, Z. Lu, X. Lin, S. Rahardja, E. P. Ong, and S. Yao, "Rate control for videophone using local perceptual cues," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 4, pp. 496–507, Apr. 2005.

[33] V. A. Mateescu, H. Hadizadeh, and I. V. Bajić, "Evaluation of several visual saliency models in terms of gaze prediction accuracy on video," in *Proc. IEEE Globecom Workshops*, Dec. 2012, pp. 1304–1308.

[34] H. Hadizadeh, I. V. Bajić, and G. Cheung, "Video error concealment using a computation-efï¬'cient low saliency prior," *IEEE Trans. Multimedia*, to appear.

[35] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami Beach, FL, USA, Jun. 2009, pp. 1597–1604.

[36] A. Torralba and A. Oliva, "Statistics of natural image categories," *Netw., Comput. Neural Syst.*, vol. 14, no. 3, pp. 391–412, 2003.

[37] H. Hadizadeh, "Visual saliency in video compression and transmission," Ph.D. dissertation, School Eng. Sci., Simon Fraser Univ., Apr. 2013.

[38] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Vis. Cognit.*, vol. 12, no. 6, pp. 1093–1123, 2005.

[39] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Amer. A*, vol. 24, no. 2, pp. B61–B69, Dec. 2007.

[40] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vis. Res.*, vol. 47, no. 19, pp. 2483–2498, Sep. 2007.

[41] Y.-M. Chen and I. V. Bajić, "Motion vector outlier rejection cascade for global motion estimation," *IEEE Signal Process. Lett.*, vol. 17, no. 2, pp. 197–200, Feb. 2010.

[42] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 9, no. 21, pp. 3888–3901, Sep. 2012.

[43] C. Chamaret, J. C. Chevet, and O. Le Meur, "Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 1077–1080.

[44] Z. G. Li, W. Gao, F. Pan, S. W. Ma, K. P. Lim, G. N. Feng, X. Lin, S. Rahardja, H. Q. Lu, and Y. Lu, "Adaptive rate control for H.264," *J. Vis. Commun. Image Represent.*, vol. 17, pp. 376–406, Apr. 2006.

[45] N. Kamaci, Y. Altunbasak, and R. M. Mersereau, "Frame bit allocation for the H.264/AVC video coder via Cauchy-density-based rate and distortion models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 8, pp. 994–1006, Aug. 2005.

[46] Z. Chen and K. N. Ngan, "Toward rate-distortion tradeoff in real-time color video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 2, pp. 158–167, Feb. 2007.

[47] C. Yeo, Y. Han Tan, Z. Li, and S. Rahardja, "Mode-dependent transforms for coding directional intra prediction residuals," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 4, pp. 545–554, Apr. 2012.

[48] I.-M. Pao and M.-T. Sun, "Modeling DCT coefficients for fast video encoding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 4, pp. 608–616, Jun. 1999.

[49] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.

[50] B. Widrow and I. Kollar, *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

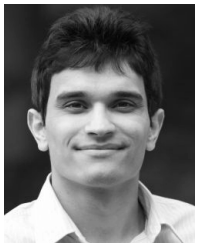
[51] H. Hadizadeh, M. J. Enriquez, and I. V. Bajić, "Eye-tracking database for a set of standard video sequences," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 898–903, Feb. 2012.

[52] *iLab Neuromorphic Vision C++ Toolkit* ver. 3.1 [Online] Available: http://ilab.usc.edu/toolkit/

[53] M. Narwaria, L. Weisi, and L. Anmin, "Low-complexity video quality assessment using temporal quality variations," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 525–535, Mar. 2012.

[54] M. Narwaria and L. Weisi, "Machine learning based modeling of spatial and temporal factors for video quality assessment," in *Proc. IEEE ICIP*, Sep. 2011, pp. 2513–2516.

[55] J. T. McClave and T. Sincich, *Statistics*, 9th ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2003.
[56] G. Bjontegaard, "Calculation of average PSNR differences between RD curves," in *13th VCEG Meeting*,VCEG-M33, Apr. 2001.
[57] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Jun. 2004.
[58] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.
[59] T.-S. Ou, Y.-H. Huang, and H. H. Chen, "SSIM-based perceptual rate control for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 5, pp. 682–691, May 2011.
[60] M. Taylor and C. Creelman, "PEST: Efficient estimates on probability functions," *J. Acoust. Soc. Amer.*, vol. 41, no. 4A, pp. 782–787, Jan. 1967.
[61] "Recommendation BT.500-8: Methodology for the subjective assessment of the quality of television pictures," ITU, San Jose, CA, USA, Tech. Rep. ITU-R BT.500-11, 1998.
[62] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. London, U.K.: Chapman & Hall, 2007.



Hadi Hadizadeh (S'09) received the B.Sc.Eng. degree in electronic engineering from the Shahrood University of Technology, Shahrood, Iran, in 2005, the M.S. degree in electrical engineering from the Iran University of Science and Technology, Tehran, Iran, in 2008, and the Ph.D. degree in engineering science from Simon Fraser University, Burnaby, BC, Canada, in 2013.

His current research interests include perceptual image/video coding, visual attention modeling, error resilient video transmission, image/video processing, computer vision, data mining, and machine learning. He was a recipient of the Best Paper Runner-up Award at ICME 2012 in Melbourne, Australia and the Microsoft Research and Canon Information Systems Research Australia Student Travel Grant for ICME 2012. He is currently a Research Associate with Simon Fraser University, Burnaby, BC, Canada and serving as the Vice Chair of the Vancouver Chapter of the IEEE Signal Processing Society.

Ivan V. Bajić (S'99–M'04–SM'11) received the B.Sc.Eng. degree (*summa cum laude*) in electronic engineering from the University of Natal, Durban, South Africa, in 1998, the M.S. degree in electrical engineering, the M.S. degree in mathematics, and the Ph.D. degree in electrical engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 2000, 2002, and 2003, respectively.

He is currently an Associate Professor of Engineering Science at Simon Fraser University, Burnaby, BC, Canada. His current research interests include signal, image, and video processing and compression, perceptual aspects of information processing, and multimedia communications. He has authored over a dozen and co-authored another six dozen publications in these fields. He has served on the program committees of various conferences in the field, including GLOBECOM, ICC, ICME, and ICIP, was the chair of the Media Streaming Interest Group of the IEEE Multimedia Communications Technical Committee from 2010 to 2012, and is currently serving as the chair of the Vancouver Chapter of the IEEE Signal Processing Society. He was a recipient of the Skye Award and the Altech Award from the University of Natal, and the South African NRF Scholarship, all in 1998, a recipient of the IBM Research Student Travel Grant for ICIP 2003, a recipient of the SFU Endowed Research Fellowship in 2005, a recipient of the Quality Reviewer Award at ICME 2011, Best Reviewer Award at VCIP 2012, and a co-recipient of the Best Paper Runner-up Award at ICME 2012.