Video Error Concealment Using a Computation-Efficient Low Saliency Prior

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

***Abstract*—Error concealment in packet-loss-corrupted streaming video is inherently an under-determined problem, as there are insufficient number of well-defined criteria to recover the missing blocks perfectly. When a Region-of-Interest (ROI) based unequal error protection (UEP) scheme is deployed during video streaming—i.e., more visually salient regions are strongly protected—a lost block is likely to be of low saliency in the original frame. In this paper, we propose to add a low-saliency prior to the error concealment problem as a regularization term. It serves two purposes. First, in ROI-based UEP video streaming, low-saliency prior provides the correct side information for the client to identify the correct replacement blocks for concealment. Second, in the event that a perfectly matched block cannot be unambiguously identified, the low-saliency prior reduces viewer's visual attention on the loss-stricken region, resulting in higher overall subjective quality. We study the effectiveness of a low-saliency prior in the context of a previously proposed RECAP error concealment system. RECAP transmits a low-resolution (LR) version of an image alongside the original high-resolution (HR) version, so that if blocks in the HR version are lost, the correctly-received LR version can serve as a template for matching of suitable replacement blocks from a previously correctly-decoded HR frame. We add a low-saliency prior to the block identification process, so that only replacement candidate blocks with good match and low saliency can be selected. Further, we develop a low-complexity convex approximation to the well known Itti-Koch-Niebur saliency model, which enables the low-saliency error concealment problem to be solved efficiently. Experimental results show that: i) PSNR of the error-concealed frames can be increased dramatically (up to 3.6 dB over the original RECAP), showing the effectiveness of a low-saliency prior in the under-determined error concealment problem; and ii) subjective quality of the repaired video using our proposal, as confirmed by an extensive user study, is better than the original RECAP.**

***Index Terms*—Visual communication, Video signal processing, Streaming media, Inverse problems.**

I. Introduction

DESPITE ongoing efforts to further advance communication technologies, high quality real-time video streaming over best-effort, packet-switched networks remains challenging for a number of reasons. First, consumer demand for interactive streaming video (e.g., conference video such as Skype, Google Talk, etc.) continues to outpace the rate of increase in network bandwidth [1], resulting in congestion and packet queue overflows in packet-switched networks. Second, when packet losses do occur, persistent server-client retransmission is not practical due to playback constraints—a video packet arriving at decoder past its playback deadline is essentially useless. Third, new media types such as ultra-high-resolution video and multiple-view video [2] that promise enhancement of viewing experience are also further straining resource-limited networks due to their large sizes. Under these practical constraints, it is difficult to guarantee error-free delivery of the entire video from sender to receiver in a timely manner.

Given packet losses are unavoidable, many previous works [3]–[5] employed the pro-active methodology of unequal error protection (UEP) of video data, where important packets are protected more heavily, for example, using stronger Forward Error Correction (FEC) codes. Typically, more important packets contain viewer's probable Regions-of-Interest (ROI) [6] in a video frame, or regions with higher *visual saliency* [7]—where viewers most likely will focus their visual attention. In such a scheme, when a packet is lost, the affected region is very likely to be of low visual saliency. While the loss of high-saliency information is still possible, this is a rare event compared to the loss of low-saliency information. Instead of proactive error protection schemes like UEP, in this paper, we study the complementary problem of *error concealment*: given the occasional packet loss during network transmission, causing the loss of a group of macroblocks (MB) in a video frame, how to best conceal the effect of data loss at the decoder to minimize visual distortion.

According to the Oxford English Dictionary [8], to "conceal" means to keep from sight; hide; keep (something) secret; prevent from being known or noticed. Error concealment is typically an under-determined problem: there are insufficient number of well-defined criteria, such as smoothness conditions for boundary pixels adjacent to correctly received neighboring blocks [9], to recover all missing MBs perfectly. This makes choosing the appropriate set of pixels to replace the missing blocks a technically challenging problem. In this paper, we propose to add a *low-saliency prior* to the error concealment problem as a regularization term. It serves two purposes. First, in ROI-based UEP video streaming, low-saliency prior is likely the correct side information for the lost block and helps the client identify the correct replacement block for concealment. Second, in the event that a perfectly matched block cannot be identified, the low-saliency prior reduces viewer's visual attention on the loss-stricken spatial region, resulting in higher

Manuscript received December 11, 2012; revised April 03, 2013; accepted June 04, 2013. Date of publication September 06, 2013; date of current version November 13, 2013. This work was supported in part by the NSERC Grant number RGPIN 327249. This paper is an extended version of the original paper which appeared in the Proceedings of IEEE ICME 2012 Conference and was among the top-rated 4% of ICME'12 submissions. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jianfei Cai.

H. Hadizadeh and I. V. Bajić are with the School of Engineering Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada (e-mail: hadi.sfu@gmail.com; ibajic@ensc.sfu.ca).

G. Cheung is with National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan (e-mail: cheung@nii.ac.jp).

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/TMM.2013.2281024

overall subjective quality. In a way, the low-saliency prior tries to make error concealment live up to its name by attempting to hide concealed blocks from viewers' attention.

We study the effectiveness of a low-saliency prior in the context of a previously proposed RECAP error concealment system [10]. RECAP transmits a low-resolution (LR) version of a video frame alongside the original high-resolution (HR) version, so that if blocks in the HR version are lost, the correctly-received LR version serves as a template for matching of suitable replacement blocks from a previously correctly-decoded HR frame. We add a low-saliency prior to the block identification process, so that only replacement candidate blocks with good match *and* low saliency can be selected. Further, we develop a low-complexity convex approximation to the well-known Itti-Koch-Niebur (IKN) saliency model [7], [11]. This makes it possible to formulate low-saliency error concealment as a convex optimization problem and solve it efficiently using convex optimization techniques. Indeed, the complexity of the proposed method can be orders of magnitude lower compared to our previous concealment method in [12], while the resulting video quality is equal or better. Specifically, experimental results show that: i) PSNR of the error-concealed frames can be increased dramatically—up to 3.6 dB over the original RECAP, and up to 0.7 dB compared to our earlier method in [12], showing the effectiveness of a low-saliency prior in the under-determined error concealment problem; and ii) subjective quality of the repaired video using our proposal, as confirmed by an extensive user study, is better than the original RECAP.

The outline of the paper is as follows. We first discuss related work in Section II. We then present an overview of the RECAP video transmission system and the IKN saliency model as well as our earlier method in [12] in Sections II-A, II-B, and II-C, respectively. The proposed error concealment strategy with low-saliency prior is discussed in Section III-A, and the convex approximation to IKN saliency is presented in Section III-B. Finally, experimental results and conclusions are presented in Sections V and VI, respectively.

II. Related Work

In the face of challenging network conditions during real-time video streaming, UEP strategies [3]–[5] protect visually important (salient) regions of interest (ROI) more heavily. An often overlooked question in these works is how to conceal missing blocks in the less important non-ROI regions when packet losses do occur? If concealment is done in a *saliency-myopic* way, so that the resulting salient features draw unwanted attention to the (likely) imperfectly recovered non-ROI blocks, it will adversely affect the subjective visual quality. This is one of the main reasons why we apply the low-saliency prior to the error concealment problem, so that concealment in non-ROI blocks can be done in a *saliency-cognizant* manner, resulting in recovered blocks that do not draw unnecessary attention.

Although we apply our low-saliency prior to the RECAP video transmission system [10] in this paper for concreteness, we believe that the low-saliency prior itself has more general applicability to other ROI-based UEP video streaming systems that may employ other error concealment tools. For example, in [9], where smoothness condition for boundary pixels is used as one condition for recovery, low saliency can be an additional requirement to further facilitate correct block recovery. Note that in our proposed method, we address packet losses in low-saliency non-ROI spatial regions because that is the *typical* case in a packet loss event. Packet loss in more heavily protected high-saliency regions, while possible, is a *rare* case, and hence will not contribute much to the average performance of the system, as long as some default concealment scheme is performed.

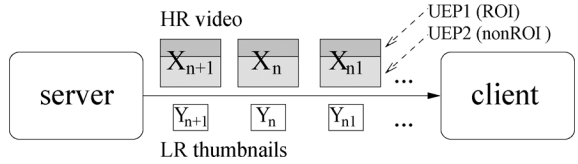


Fig. 1. Overview of RECAP packet loss recovery system.

Visual saliency—a measure of propensity for drawing visual attention—has been a subject of intense study in the past decade [7], [13], [14]. While earlier works have applied visual saliency principles to video compression [11], to the best of our knowledge, we are the first to apply saliency analysis for error concealment of streaming video. A recent evaluation of saliency models for gaze prediction in video [15] found that the well-known Itti-Koch-Niebur (IKN) saliency model [7], enhanced by the temporal features and 'FancyOne' feature integration [11], was the most accurate among the nine methods tested in that study. In our earlier work on low-saliency error concealment [12], we used the IKN model for saliency calculation. In the present paper, we provide a low-complexity convex approximation to the IKN saliency model, which allows us to formulate the error concealment problem with low-saliency prior as a convex optimization problem, leveraging existing polynomial-time convex optimization algorithms for globally optimal solutions. In so doing, as will be shown in Sections IV and V, we can find better solutions more computation-efficiently than our previous work in [12]. We believe that this convex approximation to IKN saliency can also be useful in other optimization problems that involve saliency computation.

A. Recap Video Transmission System

We first present an overview of the RECAP video transmission system [10], shown in Fig. 1, upon which we build our error concealment strategy with low-saliency prior at the decoder. Server compresses HR video into *ROI layer* and *non-ROI layer*. Using UEP, the ROI layer is more heavily protected using stronger FEC than the non-ROI layer. Typically, ROI layer contains more visually salient objects and accounts for 25% or less of the total spatial area of each frame (to be discussed in more detail in Section V). Given the relatively small size of the ROI layer, we will assume it is protected well enough that unrecoverable packet losses, as observed by the client, take place only in the non-ROI layer in the typical case.

Along with the encoded HR video, the server also low-pass filters and down-samples HR frames into LR thumbnails and transmits them with heavy protection. In practice, the size of a

thumbnail is 1/16 (down-sampled by a factor of 4 in both dimensions) of the size of the HR image, and hence it does not incur much redundant transmission overhead. While data-agnostic FEC suffers from the well-known "cliff" effect (where each block of FEC-protected source data is either recoverable in its entirety or severely damaged), thumbnail-based scheme enables a more graceful recovery, where lost HR video blocks can be partially recovered via block search in previous correctly received HR reference frame, using a LR thumbnail as template. Experimental results in [10] showed that by transmitting thumbnails, RECAP outperformed FEC-only schemes. Our goal in this paper is to improve thumbnail-based error concealment using a low-saliency prior.

B. Overview of the Visual Saliency Model

Among the existing bottom-up computational models of visual attention, the Itti-Koch-Niebur (IKN) model [7] is one of the most well-known and widely used. In this biologically plausible model, the visual saliency of different regions 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 [7].

The contrast in each feature channel is then 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 $s = c + \delta$, with $\delta \in \{3, 4\}$. 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 non-linear normalization operator to create a "conspicuity map" for each feature channel. The conspicuity maps are then 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 [11] 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 [11]. 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. A block-diagram of the IKN saliency model is shown in Fig. 2.

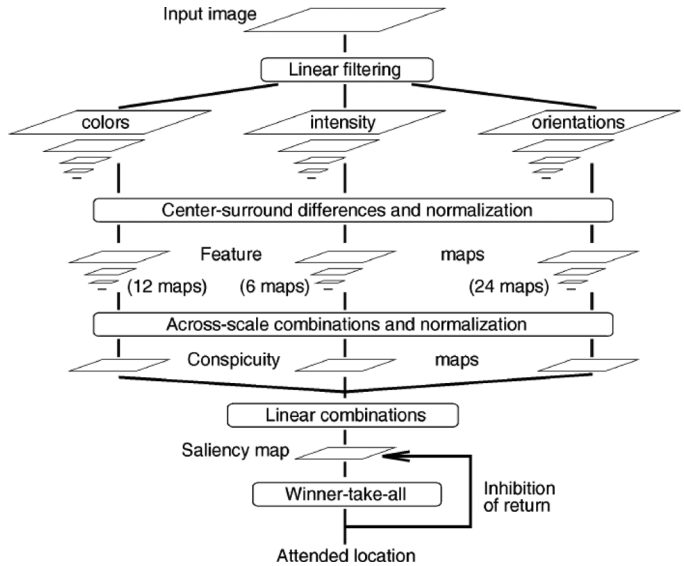


Fig. 2. A block-diagram of the IKN saliency model [7].

C. Overview of the Method From [12]

In [12], we proposed a salinecy-cognizant video error concealment method to study the effectiveness of a low-saliency prior in the context of error concealment. In that method, we added a low-saliency prior to the block identification process in RECAP, so that only replacement candidate blocks with good match and low saliency can be selected. In particular, we designed and applied four saliency reduction operators iteratively, in order to reduce the saliency of candidate blocks. These operators were: 1) Notch filter, 2) Frequency Outlier Filter, 3) Intensity and Color Contrast Reduction operator, and 4) Deblocking filter. These operators were applied on a given RECAP candidate block using the following algorithm:

1) **Step 1:** Set $j = 1$, where j refers to the index of one the four saliency reduction operators.
2) **Step 2:** Apply the j-th saliency-reduction operator on the current RECAP block.
3) **Step 3:** Project the result of Step 2 onto the thumbnail block using a project-to-thumbnail operator to make sure that the low-frequency content of the new candidate is in good match with the thumbnail block.
4) **Step 4:** Compute the saliency of the new block obtained after Step 3.
5) **Step 5:** Compute a saliency-distortion cost, where the saliency is given by Step 4, and distortion is obtained by the L_2-norm of the difference between the new candidate and the thumbnail block. If the computed cost is lower than the smallest already-known saliency-distortion cost, then go to Step 2. Otherwise go to Step 6.
6) **Step 6:** If $j < 4$, then fetch the original RECAP block again, set $j = j + 1$, and go to Step 2. Otherwise end.

The above algorithm was performed on the best K RECAP candidates whose L_2-difference with respect to the thumbnail block is the lowest. In the end, the reconstructed block whose saliency-distortion cost was the lowest was chosen as the final replacement block. To compute the saliency of the new block in Step 4, the IKN saliency model from Section II-B was utilized.

In this paper, we extend our earlier work [12] in two ways. First, the objective function in the present paper is somewhat improved. In the new objective function, we allow two different weights for the matching to the thumbnail and matching to RECAP candidates. This enables higher emphasis on low-frequency matching to the thumbnail block, which is more reliable

than RECAP candidate blocks. And second, the saliency estimation operator in the new approach is improved in terms of computational complexity. Instead of the IKN model, we develop a convex approximation to the IKN saliency, which allows us to make the objective function convex. With the help of this approximation, we are able to solve the error concealment problem at a significantly lower computational cost.

III. The Proposed Method

In this section, we present our proposed video error concealment method. In the sequel, capital bold letters (e.g., $\mathbf{X}$) denote matrices, lowercase bold letters (e.g., $\mathbf{x}$) denote vectors, and italic letters (e.g., x or X) represent scalars.

A. Problem Formulation

Consider a video frame $\mathbf{F}$ in which some blocks from non-ROI (i.e., low-salient) regions have been lost. Let $\mathbf{X}$ be a lost block of size $m \times m$, and $\mathcal{N}(\mathbf{X})$ be a $w \times h$ window in $\mathbf{F}$, with $w, h \geq m$, such that it covers only the available blocks in $\mathbf{F}$ (i.e., correctly-decoded or already-concealed blocks) in the neighborhood of $\mathbf{X}$, as well as the location of $\mathbf{X}$ itself. Let $\mathcal{S}(\mathcal{N}(\mathbf{X}))$ be a saliency operator that computes the saliency of block $\mathbf{X}$ within $\mathcal{N}(\mathbf{X})$. Also, let $vec(\mathbf{X})$ be the vectorization operator that vectorizes its input matrix $\mathbf{X}$ in a raster scan, $\mathbf{D}$ be a down-sampling matrix [16], $\mathbf{L}$ be a low-pass FIR filter [16], and $\widetilde{\mathbf{L}}$ be the high-pass FIR complement of $\mathbf{L}$, i.e., $\widetilde{\mathbf{L}} = \mathbf{I} - \mathbf{L}$, where $\mathbf{I}$ is the identity matrix.

Our goal is to reconstruct the missing block $\mathbf{X}$ so that the reconstructed block, $\widehat{\mathbf{X}}$, has low saliency after reconstruction. To achieve this goal, we propose the following algorithm, which is applied on every lost block in $\mathbf{F}$ in a raster-scan order:

- **Step 1:** Apply the RECAP algorithm on the missing block $\mathbf{X}$ to obtain the best K RECAP candidates $\mathbf{R}_k$, $k = 1, \ldots, K$, whose L_2 difference with respect to the thumbnail block $\mathbf{T}$ is the lowest.
- **Step 2:** Compute the 2-D DCT of all available spatial neighbors of $\mathbf{X}$. Let $\mathbf{B}_l$ be a matrix whose elements are set to the minimum DCT coefficients of the available spatial neighbors of $\mathbf{X}$. Similarly, let $\mathbf{B}_u$ be a matrix whose elements are set to the maximum DCT coefficients of the available spatial neighbors of $\mathbf{X}$.
- **Step 3:** Given a RECAP candidate block $\mathbf{R}_k$, solve the following minimization problem to obtain the reconstructed block $\widehat{\mathbf{X}}_k$

$$\begin{aligned}\widehat{\mathbf{X}}_k = \underset{\mathbf{X}}{\operatorname{argmin}} \Big[& \mathcal{S}\left(\mathcal{N}(\mathbf{X})\right)] \\ & + \lambda_1 \left\| \mathbf{D}\mathbf{L} vec(\mathbf{X}) - vec(\mathbf{T}) \right\|_2^2 \\ & + \lambda_2 \left\| \widetilde{\mathbf{L}} vec(\mathbf{X}) - \widetilde{\mathbf{L}} vec(\mathbf{R}_k) \right\|_2^2 \Big], \quad (1) \\ \text{subject to } & \mathbf{B}_l \leq \mathbf{\Phi} \mathbf{X} \mathbf{\Phi}^t \leq \mathbf{B}_u,\end{aligned}$$

  where λ_1 and λ_2 are two positive real scalars, $\|.\|_2$ denotes the L_2-norm, and $\mathbf{\Phi}$ is the 2-D DCT matrix, which is of the same size as $\mathbf{X}$. The down-sampling factor of $\mathbf{D}$ is set to the same down-sampling factor that is used to generate $\mathbf{T}$, and $\mathbf{L}$ is used to avoid aliasing due to down-sampling.
- **Step 4:** Repeat Step 3 for all the K RECAP candidates. Select the candidate with the smallest objective function value (1) as the final reconstructed block $\widehat{\mathbf{X}}$.

The first term in the objective function in (1) measures the saliency of the reconstructed block within $\mathcal{N}(\mathbf{X})$. The minimization of this term ensures that the reconstructed block has low saliency after reconstruction. At the same time, the constraint defined in (1) tries to eliminate any potential frequency outliers in the reconstructed block by restricting the frequency content of the reconstructed block to be within the extremes of the frequency content of its available neighboring blocks. In [12], this was one of the approaches for limiting the saliency of the reconstructed block.

The second term of the objective function in (1) ensures that the reconstructed block remains in good match with the thumbnail block, while the third term in (1) tries to match the high-frequency content in the candidate block to that of the RECAP candidate block. In practice, λ_1 should be set to a larger value than λ_2. The reason is that the thumbnail block can be considered as a very reliable side information for the low frequency content of the missing block, so it makes sense to enforce a very good match to the thumbnail block, i.e., large λ_1. However, the match does not have to be exact, because the thumbnail block has been quantized and compressed as well, and we do not want to over-fit the low frequency content of the reconstructed block to the quantized thumbnail.

Unlike the low-frequency content in the second term in (1), we do not have a very reliable side information for the high-frequency content of the missing block in the third term in (1). All we know comes from the high frequency information of the RECAP candidate block, which might not be the same as the original high frequency content of the missing block. Hence, λ_2 should be set to a smaller value than λ_1. We found experimentally that $\lambda_1 = 1.5$ and $\lambda_2 = 0.5$ work well, so we used these values in all the simulations in Section V. The hope is that saliency consideration will provide sufficient additional information to reconstruct the high-frequency content of the missing block reasonably well.

B. The Saliency Operator $\mathcal{S}(\mathcal{N}(\mathbf{X}))$

The error concealment problem formulation in Section III-A involves the saliency operator $\mathcal{S}(\mathcal{N}(\mathbf{X}))$ that computes the saliency of $\mathbf{X}$ within $\mathcal{N}(\mathbf{X})$. In our previous work [12], we used the IKN saliency model described in Section II-B as an implementation of $\mathcal{S}(\mathcal{N}(\mathbf{X}))$. However, this approach has two disadvantages: (i) it is computationally expensive, as discussed in [17], and (ii) it is non-convex in $\mathbf{X}$, making it difficult to find the globally optimal solution to (1). In this section we propose a solution to these problems by introducing a convex approximation to the IKN saliency. With a saliency operator $\mathcal{S}(\mathcal{N}(\mathbf{X}))$ that is convex in $\mathbf{X}$, the optimization problem in (1) becomes convex (the last two terms in the objective function are already convex, as is the constraint), making it possible to solve (1) using a variety methods for convex optimization [18], [19]. We will demonstrate in Section V that our convex saliency operator approximates the IKN saliency very well, yet has an advantage of being simpler to compute and easy to integrate into various optimization problems.

1) A Convex Approximation to IKN Saliency: Our convex approximation to the IKN saliency consists of two parts: spatial and temporal. We will first show how to compute saliency for a given block $\mathbf{X}$, that is $\mathcal{S}(\mathbf{X})$, and then how to compute the saliency of $\mathbf{X}$ within a neighborhood, that is $\mathcal{S}(\mathcal{N}(\mathbf{X}))$.

The dyadic Gaussian pyramid employed in the IKN model approximately halves the normalized frequency spectrum of the input image at each level due to the successive low-pass filtering. Since the normalized frequency of the original image at level 0 is $[0, \pi]$, the normalized frequency spectrum at level c of the pyramid is in the range $[0, \pi/2^c]$. For instance, the normalized frequency spectrum at levels 4 and 8 will be, respectively, in the range $[0, \pi/16]$ and $[0, \pi/256]$. As mentioned in Section II-B, in the IKN model, a center-surround feature map at center level $c \in \{2, 3, 4\}$ and surround level $s = c + \delta$, with $\delta \in \{3, 4\}$, is computed by interpolating the surround level to the center level followed by point-by-point subtraction. Hence, the normalized frequency spectrum of the center-surround feature map at center level c and surround level s will be in the range $[\pi/2^s, \pi/2^c]$. To compute the conspicuity map of each feature channel, all the computed center-surround feature maps are resized to the size of level 4. Hence, the upper limit of the normalized frequency spectrum of the obtained conspicuity map is capped by $\pi/16$. Since the smallest surround map is at level 8, we conclude that the IKN model uses the image content in the normalized frequency range $[\pi/256, \pi/16]$ to construct the saliency map, as already observed in [12], [14]. Note that the normalized frequency here is defined with respect to the original image regardless of the resolution.

To compute spatial saliency, we need a way to use the pixels of a given block $\mathbf{X}$ to estimate the saliency of the original image at that position. Based on the discussion above, it seems natural to try to recapture the portion of the image signal from the normalized frequency range $[\pi/256, \pi/16]$ at the position of the block $\mathbf{X}$. However, 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 one examines the spectrum of the block $\mathbf{X}$.

In order to address this issue we take the following approach. Consider the original image spectrum in the normalized frequency range $[0, \pi]$. We think of the image signal in the normalized frequency range $[\pi/256, \pi/16]$ as the "signal," and the signal in the remaining part of the spectrum, $[0, \pi/256) \cup (\pi/16, \pi]$, as "noise." After extracting a block from the image, both the signal and the noise leak from their native frequency bands into other bands. The spectrum of the block $\mathbf{X}$ is the sum of the leaked spectra of the signal and the noise. We need to extract the signal from noise. Since the signal and the noise come from non-overlapping frequency bands in the original image, they are orthogonal. Wiener filter is the optimum linear filter for extracting the signal from noise, and when the signal and noise are orthogonal, its transfer function is [20]

$$H(\omega) = \frac{S_S(\omega)}{S_S(\omega) + S_V(\omega)}, \tag{2}$$

where $S_S(\omega)$ is the power spectral density of the signal, and $S_V(\omega)$ is the power spectral density of the noise. Hence, Wiener filter is a frequency-domain weighting function [21].

We perform Wiener filtering in the DCT domain, rather than DFT domain, because DCT is simpler to compute (no need for complex arithmetic) and its efficient implementations are readily available in various image and video codecs. Let $\mathbf{Z}_\mathbf{X}(j, l)$ be the (j, l)-th 2-D DCT coefficient of $\mathbf{X}$. The Wiener-filtered coefficient is

$$\mathbf{Z}^W_\mathbf{X}(j, l) = \mathbf{H}(j, l)\mathbf{Z}_\mathbf{X}(j, l), \tag{3}$$

where $\mathbf{H}(j, l)$ is a coefficient that should be computed as in (2) based on signal and noise powers at the (j, l)-th 2-D DCT coefficient. A common way to design a Wiener filter is to postulate certain signal and noise models, and derive the filter from the resulting power spectral densities [22]. We use the $1/f$-model, which is thought to be an excellent model for natural images [23], as a starting point; our "signal" is the part of the $1/f$ signal in the frequency band $[\pi/256, \pi/16]$, and our "noise" is the part of the $1/f$ signal in the remainder of the spectrum.

To compute $\mathbf{H}(j, l)$, we proceed as follows. We generate a deterministic $1/f$ 2-D signal that covers the frequency band $[\pi/256, \pi/16]$, at a size equal to the target image resolution. We then extract from this signal a block whose size is equal to the block size of interest and perform a 2-D DCT on it. Let us denote the resulting DCT by $\mathbf{Z_S}(i, j)$. Then $\mathbf{Z_S}^2(i, j)$ is the signal power associated with that DCT coefficient, corresponding to $S_S(\omega)$ in (2). Similarly, we find the noise power associated with DCT coefficient (i, j), $\mathbf{Z_V}^2(i, j)$ by using a deterministic $1/f$ 2-D signal that covers the frequency band $[0, \pi/256) \cup (\pi/16, \pi]$. The DCT-domain Wiener filter coefficients are then given by

$$\mathbf{H}(j, l) = \frac{\mathbf{Z_S}^2(j, l)}{\mathbf{Z_S}^2(j, l) + \mathbf{Z_V}^2(j, l)}. \tag{4}$$

Note that $\mathbf{H}(j, l)$ depends on image resolution and the block size, due to the way $\mathbf{Z_S}(i, j)$ and $\mathbf{Z_V}(i, j)$ are computed, but can be easily pre-computed for typical resolutions and block sizes. Fig. 3 shows the Wiener coefficients obtained by the proposed method for two standard resolutions, 352×288 and 1024×768, and a block size of 16×16. Observe that low-frequency coefficients are higher than high-frequency coefficients, as one would expect for a signal that came from the normalized frequency band $[\pi/256, \pi/16]$ in the original image. However, due to spectral leakage, some of the higher frequency coefficients also contain part of the signal, which makes their coefficients non-zero.

Our approximation to the spatial saliency of block $\mathbf{X}$ is the power of the Wiener-filtered signal $\mathbf{Z}^W_\mathbf{X}$, that is

$$\mathcal{S}_{spatial}(\mathbf{X}) = \sum_{(j,l)} (\mathbf{Z}^W_\mathbf{X}(j, l))^2 = \sum_{(j,l)} \mathbf{H}^2(j, l)\mathbf{Z_X}^2(j, l). \tag{5}$$

If block $\mathbf{X}$ has multiple color channels (e.g., YUV), the power in all channels is added together. Since DCT is a linear operation, as is Wiener filtering, while squaring is a convex operation, the saliency estimate $\mathcal{S}_{spatial}(\mathbf{X})$ in (5) is convex in $\mathbf{X}$.

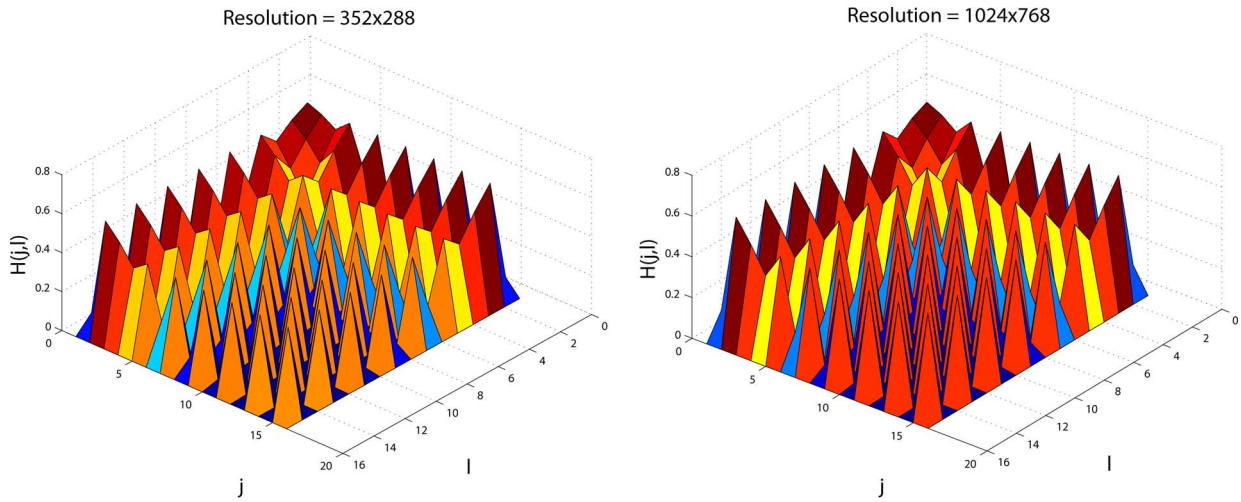


Fig. 3. Wiener coefficients for a 16 × 16 block for two common resolutions.

Next, we provide a convex approximation to temporal saliency. Let $\mathbf{X}_0$ be the co-located block of $\mathbf{X}$ in the previous frame, and let $\mathbf{Q} = |\mathbf{X} - \mathbf{X}_0|$ be the residual block obtained by taking the absolute difference between $\mathbf{X}$ and $\mathbf{X}_0$. Our approximation to the temporal saliency of block $\mathbf{X}$ is the power of the Wiener-filtered signal $\mathbf{Z}_\mathbf{Q}^W$, that is

$$\mathcal{S}_{temporal}(\mathbf{X}) = \sum_{(j,l)} (\mathbf{Z}_\mathbf{Q}^W(j,l))^2 = \sum_{(j,l)} \mathbf{H}^2(j,l) {\mathbf{Z}_\mathbf{Q}}^2(j,l), \tag{6}$$

where ${\mathbf{Z}_\mathbf{Q}}^2(j,l)$ is the (j,l)-th 2-D DCT coefficient of $\mathbf{Q}$. Note that $\mathcal{S}_{temporal}(\mathbf{X})$ is convex in $\mathbf{X}$ because $\mathbf{Q}$ is convex in $\mathbf{X}$, DCT and Wiener filtering are linear, and squaring is a convex operation.

In order to get the final saliency estimate of $\mathbf{X}$, we combine the spatial and temporal saliency terms as follows

$$\mathcal{S}(\mathbf{X}) = \mathcal{S}_{spatial}(\mathbf{X}) + \alpha \mathcal{S}_{temporal}(\mathbf{X}), \tag{7}$$

where α is a positive control parameter that trades off between the two saliency terms. We note that $\mathcal{S}(\mathbf{X})$ is convex in $\mathbf{X}$ because it is a non-negative linear combination of convex terms.

Up to this point, we have defined a convex saliency operator $\mathcal{S}(\mathbf{X})$ to approximate IKN saliency of $\mathbf{X}$ itself. We next define the operator $\mathcal{S}(\mathcal{N}(\mathbf{X}))$ that computes the saliency of $\mathbf{X}$ within a neighborhood $\mathcal{N}(\mathbf{X})$.

2) The Definition of $\mathcal{S}(\mathcal{N}(\mathbf{X}))$: Let $\mathcal{N}(\mathbf{X})$ be a $p \times p$ matrix of pixels in $\mathbf{F}$ (with $p > m$) such that it covers both the $m \times m$ missing block $\mathbf{X}$ and parts of the available 8-connected spatial neighbors of $\mathbf{X}$. Hence, the position of $\mathcal{N}(\mathbf{X})$ relative to $\mathbf{X}$ depends on the available neighbors of $\mathbf{X}$. In Appendix A, we describe various possible cases for defining $\mathcal{N}(\mathbf{X})$ relative to $\mathbf{X}$. The saliency $\mathcal{S}(\mathcal{N}(\mathbf{X}))$ is computed as in (7), with $\mathbf{X}$ replaced by $\mathcal{N}(\mathbf{X})$. Below we show that both the spatial and temporal saliency terms are sill convex in $\mathbf{X}$ when $\mathbf{X}$ is replaced by $\mathcal{N}(\mathbf{X})$.

Let $\mathbf{B}$ be a $p \times p$ matrix whose elements are all equal to the elements of $\mathcal{N}(\mathbf{X})$ except for the elements whose coordinates coincide with $\mathbf{X}$, which are set to zero. In other words, $\mathbf{B}$ is a matrix that contains the boundary pixels of $\mathcal{N}(\mathbf{X})$ around the missing block $\mathbf{X}$. Fig. 4 illustrates $\mathbf{X}$, $\mathcal{N}(\mathbf{X})$, and $\mathbf{B}$.

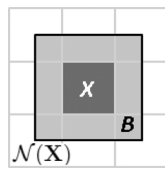


Fig. 4. An illustration of the missing block $\mathbf{X}$, and matrices $\mathcal{N}(\mathbf{X})$ and $\mathbf{B}$. Note that $\mathcal{N}(\mathbf{X})$ covers the missing block $\mathbf{X}$ and parts of the available spatial neighbors of $\mathbf{X}$. $\mathbf{B}$ is a matrix that contains the boundary pixels of $\mathcal{N}(\mathbf{X})$ around the missing block $\mathbf{X}$ (light-shaded area around $\mathbf{X}$ in this figure). Those elements of $\mathbf{B}$ whose coordinates coincide with $\mathbf{X}$ are set to zero. The saliency of $\mathbf{X}$ is computed within the area covered by $\mathcal{N}(\mathbf{X})$. In this example, it is assumed that all the 8-connected spatial neighbors of $\mathbf{X}$ are available. Depending on the availability of the spatial neighbors of $\mathbf{X}$, the area covered by $\mathcal{N}(\mathbf{X})$ changes, as discussed in Appendix A.

$\mathcal{N}(\mathbf{X})$ can be obtained by zero-padding (expanding) $\mathbf{X}$ via a matrix expansion operator, $\mathcal{Z}(\mathbf{X}, \mathcal{N}(\mathbf{X}))$, and adding the resulting matrix to $\mathbf{B}$. The matrix expansion operator, $\mathcal{Z}(\mathbf{X}, \mathcal{N}(\mathbf{X}))$, zero-pads the $m \times m$ matrix $\mathbf{X}$ up to a $p \times p$ matrix, $\mathbf{X}_e$, and can be realized as a linear operation

$$\mathbf{X}_e = \mathcal{Z}(\mathbf{X}, \mathcal{N}(\mathbf{X})) = \mathbf{MXN}, \tag{8}$$

where $\mathbf{M}$ is a binary matrix of size $p \times m$, and $\mathbf{N}$ is a binary matrix of size $m \times p$, both of which depend on $\mathcal{N}(\mathbf{X})$. The method to derive $\mathbf{M}$ and $\mathbf{N}$ based on $\mathcal{N}(\mathbf{X})$ is given in Appendix A. Finally, since

$$\mathcal{N}(\mathbf{X}) = \mathbf{X}_e + \mathbf{B} = \mathbf{MXN} + \mathbf{B} \tag{9}$$

is an affine function of $\mathbf{X}$, it is also convex in $\mathbf{X}$. Due to this, we have that

$$\mathcal{S}_{spatial}(\mathcal{N}(\mathbf{X})) = \mathcal{S}_{spatial}(\mathbf{MXN} + \mathbf{B}), \tag{10}$$

where $\mathcal{S}_{spatial}(\cdot)$ is computed as in (5),

$$\mathcal{S}_{temporal}(\mathcal{N}(\mathbf{X})) = \mathcal{S}_{temporal}(\mathbf{MXN} + \mathbf{B}), \tag{11}$$

where $\mathcal{S}_{temporal}(\cdot)$ is computed as in (6), and

$$\mathcal{S}(\mathcal{N}(\mathbf{X})) = \mathcal{S}_{spatial}(\mathcal{N}(\mathbf{X})) + \alpha\mathcal{S}_{temporal}(\mathcal{N}(\mathbf{X})), \quad (12)$$

where α is positive as in (7), are all convex in $\mathbf{X}$.

C. Solving the Minimization Problem (1)

By using a saliency operator $\mathcal{S}(\mathcal{N}(\mathbf{X}))$ that is convex in $\mathbf{X}$, the optimization problem in (1) becomes convex. The objective function is the sum of three terms. The first one is convex in $\mathbf{X}$ if the convex saliency operator from Section III-B.2 is used, as discussed above. The second and third terms are compositions of vectorization (which is convex [24]), linear filtering, and the squared L_2-norm (which is also convex [19]), making them both convex in $\mathbf{X}$. Finally, the constraint is a combination of affine functions in $\mathbf{X}$, making it convex in $\mathbf{X}$. Hence, in this case, a variety of methods for convex optimization [19], such as interior-point, ellipsoid, subgradient, etc., can be used to solve (1).

IV. Computational Complexity

In Section IV-A, we analyze the computational complexity of the error concealment method proposed in Section III-A. Following that, in Section IV-B we compare this complexity against the complexity of our previous method from [12].

A. Computational Complexity of the Proposed Method

In this section, we estimate the computational cost of the error concealment method presented in Section III-A. In the proposed method, the minimization problem defined in (1) is solved for the best K RECAP candidates whose L_2 difference with respect to the thumbnail block is the lowest. In the end, the best block whose saliency-distortion cost in (1) is the lowest, is taken as the concealed block.

To estimate the computational cost, we need to answer two questions: 1) how many evaluations of the objective function in (1) are needed? and 2) how many operations are required in each evaluation of the objective function in (1)? The first question is difficult to answer in general. As mentioned in Section III-C, a convex optimization problem can be solved relatively easily (in polynomial time) by various convex optimization methods such as interior-point and ellipsoid methods [18], [19]. However, the exact number of objective function evaluations is not easily determined. In our experiments we found that usually about 8 objective function evaluations are needed to achieve an acceptable tolerance level of $\epsilon = 10^{-6}$ when solving (1) for 16×16 blocks. Hence, for the purpose of estimating complexity, we assume that the average number of objective function evaluations in (1) is $N_e = 8$.

We now compute the number of operations that are performed in one evaluation of the objective function in (1). To find the cost for the first term in (1), we need to find the cost of the saliency operator $\mathcal{S}(\mathcal{N}(\mathbf{X}))$ defined in (12), which involves in computing $\mathcal{S}_{spatial}(\mathcal{N}(\mathbf{X}))$ and $\mathcal{S}_{temporal}(\mathcal{N}(\mathbf{X}))$. The first step in computing $\mathcal{S}_{spatial}(\mathcal{N}(\mathbf{X}))$ is to construct $\mathcal{N}(\mathbf{X})$. In practice, $\mathcal{N}(\mathbf{X})$ can be constructed by copying the $N_b \times N_b$ block $\mathbf{X}$ to the zero locations of the $p \times p$ matrix $\mathbf{B}$ (i.e., locations in which the elements of $\mathbf{B}$ are zero). To copy a $N_b \times N_b$ matrix to another place in memory, we need to update the pointer address of both the source and destination locations after reading/copying each row of the matrix. To obtain the pointer address of the next row of the matrix, we first need to increase the current row number by one, and then the convert the 2-D address of the first element of next row into a linear 1-D address. This needs 3 operations (two additions and one multiplication) [25]. Hence, we consider approximately $2 \cdot 3N_b = 6N_b$ operations for copying a $N_b \times N_b$ matrix to another place in memory. Assuming that $\mathbf{B}$ is available before solving (1), copying the $\mathbf{X}$ to the zero locations of $\mathbf{B}$ needs approximately $6N_b$ operations.

The next step is to compute the 2-D DCT of $\mathcal{N}(\mathbf{X})$. Note that the multiplication of a $A \times B$ matrix by a $B \times C$ matrix requires $A \cdot C \cdot (2B-1)$ operations. Also, computing the 2-D DCT of a $p \times p$ block requires two $p \times p$ matrix multiplications. Hence, computing the 2-D DCT of $\mathcal{N}(\mathbf{X})$ requires $2p^2(2p-1)$ operations. We then need to compute the square of the Wiener-filtered coefficients. This step needs $2p^2$ operations. Finally, all the squared Wiener-filtered coefficients should be summed up together. This step needs approximately p^2 operations. Hence, computing $\mathcal{S}_{spatial}(\mathcal{N}(\mathbf{X}))$ in the luma (Y) channel of $\mathbf{X}$ requires approximately $6N_b + p^2(4p+1)$ operations. Assuming that $\mathbf{X}$ is in YUV 4:2:0 format, the total computational cost for computing $\mathcal{S}_{spatial}(\mathcal{N}(\mathbf{X}))$ will be $1.5(6N_b + p^2(4p+1))$.

To compute $\mathcal{S}_{temporal}(\mathcal{N}(\mathbf{X}))$, we first need to compute the absolute difference between $\mathcal{N}(\mathbf{X})$ and the co-located $p \times p$ block in the previous frame in the luma (Y) channel. For this purpose, we need to construct the neighborhood in the previous frame similar to $\mathcal{N}(\mathbf{X})$. This approximately needs $6N_b$ operations. Note that $\mathcal{N}(\mathbf{X})$ is already constructed when computing $\mathcal{S}_{spatial}(\mathcal{N}(\mathbf{X}))$. Thus, this step requires about $6N_b + 2p^2$ operations, where we considered two operations for computing the absolute difference between two elements of memory. We then need to compute the 2-D DCT of the obtained residual block, which requires $2p^2(2p-1)$ operations. After that we need to compute the sum of the squared Wiener-filtered coefficients of the residual block, which requires approximately $2p^2 + p^2$ operations. Hence, computing $\mathcal{S}_{temporal}(\mathcal{N}(\mathbf{X}))$ requires approximately $2p^2(2p+3)$ operations. Based on the above analysis, computing $\mathcal{S}(\mathcal{N}(\mathbf{X}))$ requires approximately $9N_b + p^2(10p + 7.5) + 2$ operations.

To find the computational cost of the second and third terms in (1), we first note that if the size of $\mathbf{X}$ in (1) is $N_b \times N_b$, then $\mathbf{L}$ and $\widetilde{\mathbf{L}}$ are both of size $N_b^2 \times N_b^2$, while $\mathbf{D}$ (the down-sampling matrix) is a $(N_b^2/d_s^2) \times N_b^2$ matrix, where d_s is the down-sampling factor that is used to generate the thumbnail block $\mathbf{T}$. To vectorize a $N_b \times N_b$ matrix, similar to the case discussed above for copying a $N_b \times N_b$ block to another location in the memory, we consider about $6N_b$ operations. Hence, to obtain $vec(\mathbf{R}_k)$ or $vec(\mathbf{X})$, we consider approximately $6N_b$ operations. Similarly, we approximate the cost for obtaining $vec(\mathbf{T})$ by $6N_b/d_s$ operations.

To compute $\mathbf{DL}vec(\mathbf{X})$ for evaluating the second term in (1), we can first compute $\mathbf{L}vec(\mathbf{X})$, which requires $N_b^2(2N_b^2 - 1)$ operations. We can then multiply $\mathbf{D}$ by the resultant $N_b^2 \times 1$ vector. This requires additional $N_b^2(2N_b^2 - 1)/d_s^2$ operations.

Hence, in total, computing $\mathbf{DL}vec(\mathbf{X})$ costs $N_b^2(2N_b^2-1)+N_b^2(2N_b^2-1)/d_s^2+6N_b$ operations.

As a simpler alternative, however, the low-pass filtering can be performed in the DCT domain. For this purpose, we first compute the 2-D DCT of $\mathbf{X}$, which needs $2N_b^2(2N_b-1)$ operations. We then zero out the desired high frequency coefficients. This process needs approximately N_b^2 operations. We then take the inverse 2-D DCT of the obtained result to get the filtered block in the pixel domain. This step needs $2N_b^2(2N_b-1)$ additional operations. Finally, we down-sample the obtained block by a down-sampling factor d_s to get a down-sampled block of the same size as $\mathbf{T}$. We consider N_b^2/d_s^2 operations for this step. Finally, the L_2-norm of the difference between the obtained low-resolution block and $\mathbf{T}$ must be calculated. This step requires approximately $3N_b^2/d_s^2$ operations. Therefore, in total, computing the second term of (1) in the luma (Y) channel requires approximately $N_b^2(8N_b-3+4/d_s^2)$ operations.

For computing the third term in (1), a similar approach can be utilized. Specifically, we take the 2-D DCT of both $\mathbf{X}$ and $\mathbf{R}_k$, and compute the L_2-norm of the difference between the high frequency coefficients of $\mathbf{X}$ and $\mathbf{R}_k$. To compute the L_2-norm, only those high frequency DCT coefficients that are zeroed out when computing the second term of (1) are utilized. Note that since DCT is a unitary transform, we do not need to take the inverse 2-D DCT to compute the L_2-norm difference in the pixel domain. The 2-D DCT of $\mathbf{X}$ is available after computing the second term in (1). Thus, we only need to compute the 2-D DCT of $\mathbf{R}_k$, which can be pre-computed before evaluating (1). Considering $3N_b^2$ operations for computing the L_2-norm, computing the third term in (1) in the luma (Y) channel requires approximately $3N_b^2$ operations.

In summary, we conclude that computing the second term in (1) for all three YUV 4:2:0 channels of $\mathbf{X}$ in each evaluation of the objective function requires approximately $1.5\cdot(N_b^2(8N_b-3+4/d_s^2)$ operations. For the third term, the cost is approximately $1.5\cdot(3N_b^2)$ operations.

To evaluate the constraint in (1), we need to compare the 2-D DCT of $\mathbf{X}$ with both $\mathbf{B}_l$ and $\mathbf{B}_u$. For these two comparisons, we consider $2N_b^2$ operations. The 2-D DCT of $\mathbf{X}$ is computed during the evaluation of the second term in (1) as described above. Hence, assuming that $\mathbf{B}_l$ and $\mathbf{B}_u$ are pre-computed before solving (1), evaluating the constraint for all three YUV 4:2:0 channels of $\mathbf{X}$ requires $1.5\cdot(2N_b^2)=3N_b^2$ operations.

Before evaluating (1), we need to compute the 2-D DCT of $\mathbf{R}_k$, which requires $1.5\cdot 2N_b^2\cdot(2N_b-1)$ operations. We also need to compute $\mathbf{B}_l$ and $\mathbf{B}_u$. Assuming the worst case (from the point of view of complexity) that all the four spatial neighbors of $\mathbf{X}$ are available, and that none of them are neighbors of any previously concealed blocks (in which case their 2-D DCT would already be available), we need to compute the 2-D DCT of all four neighbors. This requires $1.5\cdot 4\cdot 2N_b^2(2N_b-1)$ operations. Assuming that finding the minimum or maximum of 4 DCT coefficients needs 3 operations, we can estimate the cost for computing $\mathbf{B}_l$ or $\mathbf{B}_u$ as $1.5\cdot 3N_b^2=4.5N_b^2$ operations. Hence, the total cost for computing $\mathbf{R}_k$, $\mathbf{B}_l$, and $\mathbf{B}_u$ is approximately $N_b^2(30N_b-11)$ operations.

Overall, for the YUV 4:2:0 video format, the total computational cost of the proposed error concealment method for reconstructing a $N_b\times N_b$ block $\mathbf{X}$ within a $p\times p$ neighborhood $\mathcal{N}(\mathbf{X})$ is approximately

$$\begin{aligned}\zeta(PM) &\approx N_b^2(30N_b-11)+N_e\Bigg((9N_b+p^2(10p+7.5)+2)\\ &\quad+1.5\left(N_b^2\left(8N_b-3+\frac{4}{d_s^2}\right)\right)+3N_b^2\Bigg)\\ &=N_b^2(30N_b-11)+N_e\Bigg(12N_b^3-1.5N_b^2\\ &\quad+9N_b+\frac{6}{d_s^2}+10p^3+7.5p^2+2\Bigg).\end{aligned}\tag{13}$$

B. Comparison With the Method From [12]

In our experiments, the size of each missing block $\mathbf{X}$ is 16×16. Therefore, $N_b=16$. We also set $d_s=4$ and $p=20$ so that the window $\mathcal{N}(\mathbf{X})$ covers a 20×20 region. With these parameters in (13), we get

$$\zeta(PM)\approx 1175379.\tag{14}$$

Meanwhile, our previous error concealment method from [12] requires the computation of IKN saliency [11] within an adaptive window of size $W_0\times H_0$ that includes the missing block and its causal spatial neighborhood. As discussed in [17], the number of operations required to reconstruct a missing block of size $N_b\times N_b$ pixels by our previous method in [12] is

$$\zeta(OM)\approx 36.56N_b^3+\left(\log_2\frac{N_b^6}{4}+1104.5\right)N_b^2+5882\cdot W_0\cdot H_0.\tag{15}$$

Substituting $N_b=16$ we obtain

$$\zeta(OM)\approx 438133+5882\cdot W_0\cdot H_0.\tag{16}$$

Note that W_0 and H_0 can be as small as N_b, the size of the block, or as large as W and H, the width and height of the frame. The number of operations involved in reconstructing a block varies depending on the position of that block, which determines W_0 and H_0 [12]. At the low end, when $W_0=H_0=N_b=16$, $\zeta(PM)\approx 0.60\cdot\zeta(OM)$, making the proposed method roughly 40% less costly than the method from [12]. At the high end, when W_0 and H_0 are equal to the dimensions of the frame, then even for a CIF resolution of 352×288 (which may be considered small by today's standards), we obtain $\zeta(PM)\approx 0.002\cdot\zeta(OM)$. Hence, in this case, the proposed method has only $1/500$-th of cost of the method from [12]. In practice, the computational savings will be somewhere between these two extreme values. To get a feeling for the average case, consider CIF resolution video (352×288). Assuming that each block is equally likely to be damaged, the expected (average) position of the damaged block is at the center of the frame, and the expected values of W_0 and H_0 are $352/2=176$ and $288/2=144$, respectively. Using these values in (16) and comparing the result with $\zeta(PM)$, we find $\zeta(PM)\approx 0.008\cdot\zeta(OM)$. That is, the

expected cost of the proposed method in the case of CIF resolution video is about 1/120-th of that in [12].

V. Experimental Results

In this section, we first assess how close the convex approximation from Section III-B is to the actual IKN saliency. After that, we evaluate the performance of the saliency-cognizant error concealment method from Section III-A by comparing it with the original RECAP algorithm, as well as our previous error concealment method from [12].

A. Assessment of the Proposed Convex Approximation to IKN Saliency

As explained in Section III-B, our approximation to IKN saliency has two terms: spatial and temporal. The approximation accuracy of the spatial term (5) with block size 16×16 is first assessed on two popular still image datasets with associated ground truth eye-tracking data (fixation points). The first dataset is the so-called Toronto data set [26], which contains 120 RGB images (688×512 pixels) of outdoor and indoor scenes with eye-tracking data of 20 subjects. The eye-tracking data of this dataset was recorded in a free-viewing task at a viewing distance of 75 cm, and each image was presented for 4 seconds with a 2-second gray mask in between. The second data set is the so-called MIT data set [27], [28], which contains 1003 RGB indoor and outdoor images (1024×768 pixels) with eye-tracking data of 15 subjects. In this dataset, there were 779 landscape images and 228 portrait images. The viewing distance was fixed at 48 cm, and each image was displayed for 3 seconds.

The accuracy of saliency detection is measured by the well-known receiver operating characteristic (ROC) area under curve (AUC) measure [28]–[30]. In order to compute the AUC score for a saliency map, the hit rate is computed by determining the locations where the saliency map is above a threshold and a fixation is present in those regions. Similarly, the false alarm rate is computed by finding the locations where the saliency values are above the threshold while there is no fixation present in those regions. The ROC curve is then generated by varying the threshold to cover a wide range of possible saliency values. The area under the ROC curve is then considered as the AUC score. An AUC value of 0.5 corresponds to pure chance, a value greater than 0.5 indicates positive correlation, and 1.0 corresponds to a perfect prediction of eye fixations [29].

Table I shows the average AUC scores of the spatial IKN model and our approximation on each of the two datasets. As seen from the table, the average AUC scores of the proposed approximation are very close to the average AUC scores of the IKN model in each of the two datasets, indicating good approximation. To check for the statistical significance of this observation, we performed a paired t-test [31] between the AUC scores on each pair of images in the two datasets, with the null hypothesis that the two samples come from distributions with equal means and unknown variances. The resultant p-values [31] are also reported in Table I. In experimental sciences, as a rule of thumb, the null hypothesis is rejected when $p < 0.05$. As seen from Table I, the p-value for both data sets is well above 0.05, which indicates that the two sets of AUC scores are statistically very similar, i.e., virtually indistinguishable.

TABLE I
Average AUC Scores of the Spatial IKN Saliency and the Proposed Approximation on Two Common Datasets

Dataset	IKN Saliency Model	Proposed Approximation	p-value
Toronto	0.6512	0.6468	0.6233
MIT	0.6261	0.6244	0.6426

To further compare the saliency maps produced by the proposed spatial approximation (5) with those produced by the original IKN model, we employed the Kullback–Leibler Divergence (KLD) [32]. For this purpose, we first normalized each saliency map, and then considered the saliency map as a 2-D probability distribution. We then computed the average symmetric KLD [32] between the two sets of saliency maps on both datasets. The average symmetric KLD on the Toronto data set was 0.01630, and it was 0.01355 on the MIT data set. Averaging these two, taking into account the number of images in each, the overall average KLD between the IKN saliency maps and our approximation was 0.0138.

In order to get a feeling for what KLD of 0.0138 between saliency maps means, we performed an experiment using JPEG coding and compared IKN saliency maps of original and encoded images. For this purpose, we compressed the images in the two datasets with a JPEG encoder at various quality factors, and for each quality factor, we computed the average symmetric KLD between the IKN saliency maps of the original images and the IKN saliency maps of the compressed images. We also computed the average PSNR for each quality factor. We then repeated this experiment until we got an average symmetric KLD of 0.0138. At this KLD, the average PSNR was about 40.2 dB. Therefore, one can say that the loss in accuracy in our approximation for spatial IKN saliency is comparable to that incurred in high-quality image compression that results in a PSNR of about 40.2 dB. Fig. 5 shows several sample images from the Toronto data set along with their IKN saliency maps as well as the saliency maps generated by the proposed spatial saliency approximation.

As a further illustration, we repeated the above experiment with a "naive" spatial saliency approximation that uses only five DCT coefficients $(j,l) \in \{(0,1),(0,2),(1,1),(1,0),(2,0)\}$ and sets their weight to 1 in (5), while setting the weight of other coefficients to zero. These coefficients correspond to the normalized frequency band $[\pi/256, \pi/16]$ of the 16×16 block. As shown in Fig. 3, these coefficients do end up with some of the highest Wiener weights, but this approach ignores spectral leakage, which is why we call it "naive." This method produces saliency maps with an average KLD of 0.0165 with respect to IKN maps, over the two datasets. Using the JPEG coding analogy above, the average KLD of 0.0165 corresponds to compression at 38.5 dB. Hence, although not as good as the Wiener-based approach, this "naive" method still performs reasonably well in terms of spatial saliency approximation.

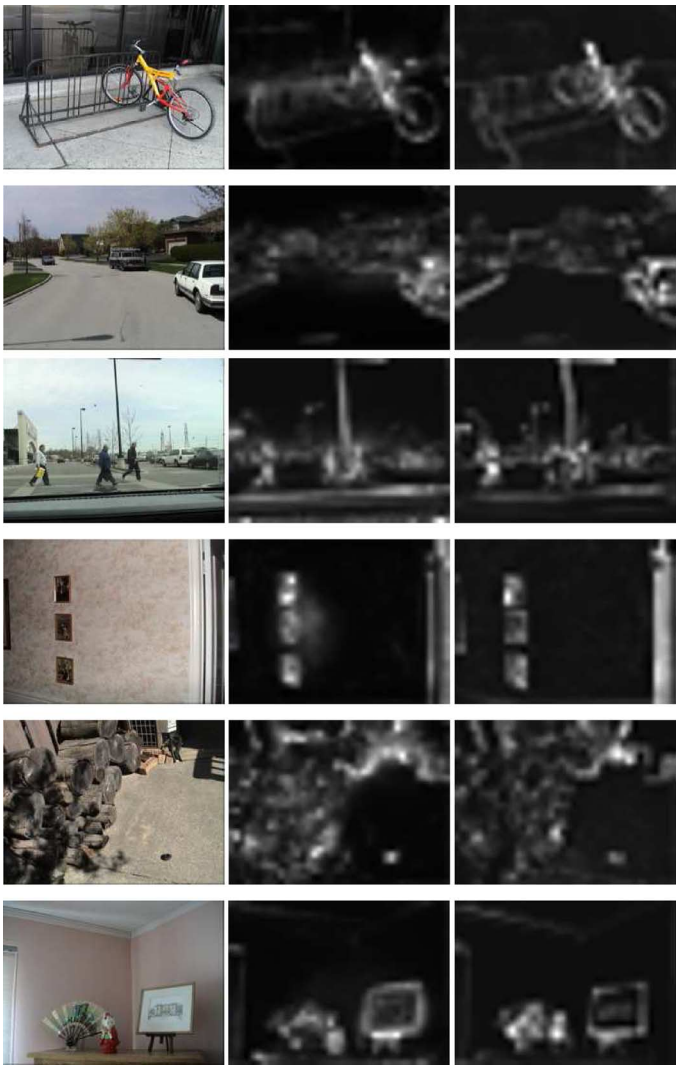


Fig. 5. Sample images from the Toronto data set (left) along with their IKN saliency map (middle) and the saliency map generated by the proposed approximation (right).

We next assess the temporal saliency approximation together with spatial saliency approximation in the context of saliency estimation in video. For this purpose, the benchmark IKN model is outfitted with a flicker and motion channel for temporal saliency estimation. Table II compares the spatial IKN saliency with the approximation in (5), the temporal IKN saliency with the approximation in (6), as well as the full IKN saliency with the combined saliency approximation in (7) on ten standard CIF sequences at 30 frames per second (fps). A brief description of the content of these standard sequences can be found in [15]. As seen in the table, the average symmetric KLD between the spatial IKN saliency and our approximation is 0.0236, which corresponds to a PSNR of 36.2 dB using the JPEG coding analogy above. The average symmetric KLD between the temporal saliency maps is 0.0151, which corresponds to a PSNR of 39.2 dB. Finally, the average symmetric KLD between the full saliency maps is 0.0178, corresponding to a PSNR of 38.3 dB, which is thought to be a fairly decent quality. Therefore, according to the above results, we conclude that the accuracy of our proposed convex approximation to IKN saliency is quite satisfactory.

TABLE II
AVERAGE SYMMETRIC KLD BETWEEN THE IKN SALIENCY AND OUR APPROXIMATION ON TEN STANDARD CIF SEQUENCES

Sequence	Spatial Saliency	Temporal Saliency	Full Saliency
Bus	0.0163	0.0198	0.0147
City	0.0144	0.0101	0.0115
Crew	0.0107	0.0095	0.0087
Foreman	0.0304	0.0110	0.0189
Flower Garden	0.0138	0.0192	0.0117
Harbour	0.0389	0.0103	0.0243
Mobile Calendar	0.0184	0.0151	0.0146
Soccer	0.0162	0.0046	0.0106
Stefan	0.0477	0.0201	0.0301
Tempete	0.0294	0.0315	0.0333
Average	**0.0236**	**0.0151**	**0.0178**

B. Evaluation of the Proposed Error Concealment Method

In order to evaluate the performance of the proposed error concealment method, we used six standard 30 fps sequences: *Soccer* (704 × 576), *RaceHorses* (416 × 240), *Tractor* (768 × 432), *Crew* (704 × 576), *Park Joy* (1280 × 720), and *Pedestrian Area* (1280 × 720). All sequences were 250 frames long. A brief description of the six test sequences is as follows. *Soccer* is a shot of a soccer field with several persons playing soccer in it. The players frequently enter and leave the scene. The camera initially pans to the right following the players and then follows the soccer ball as it comes closer to the camera. *Race Horses* is a shot of several persons riding race horses in a field, and the camera is following the persons slowly. In sequence *Tractor*, a tractor is working in a farm field, and the camera is following the tractor. The tractor occupies a large portion of the scene. *Crew* is a shot of a crew of astronauts entering a hallway and coming towards the camera, while being flash-photographed. The camera is initially relatively static (although hand-held), and starts moving more noticeably after frame 110. In sequence *Park Joy*, a number of people are running in a park, and the camera is tracking them slowly from afar, behind the trees. The relative size of people is small compared to the background and the tree trunks in the foreground in this sequence. *Pedestrian Area* is a shot of a pedestrian area. The camera is at a low position, and people pass by very close to the camera. The camera is static in this sequence. Fig. 6 shows a thumbnail of each of the six test sequences.

In our experiments, *RaceHorses* was encoded at 700 kbps, *Soccer*, *Tractor*, and *Crew* were encoded at 1400 kbps, and the other two sequences were encoded at 5800 kbps using the H.264/AVC JM 18.0 reference software [33], with the GOP structure IPPP. The thumbnail videos were created by downsampling their corresponding high resolution (HR) videos by a factor of 4 in each dimension, and were encoded at 10% of the bitrate of their HR version, using the same encoder structure as their HR version. We set $m = 16$, $p = 20$, and $\alpha = 1$.

In order to find the most salient regions or ROIs, we first computed the full IKN saliency map of each video frame of each sequence. The saliency map of each frame was then binarized based on the 75-th percentile of the saliency map of that frame.

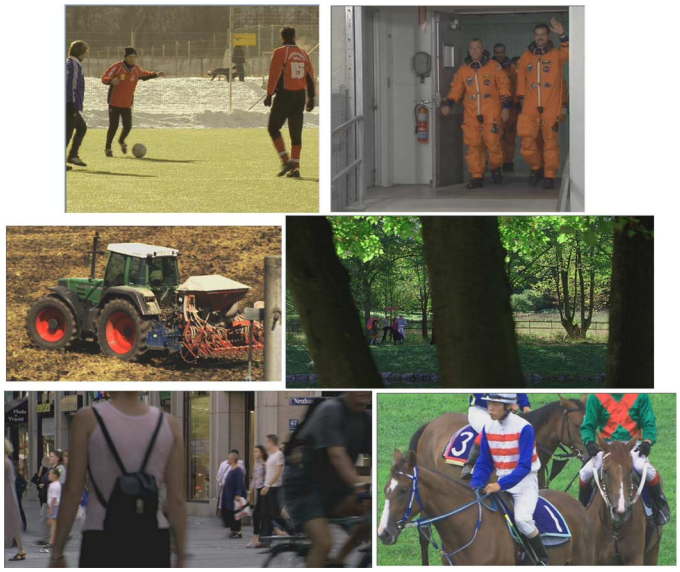


Fig. 6. Sample frames of the test sequences used in the experiments. Top row from left to right: *Soccer* and *Crew*. Middle row from left to right: *Tractor* and *Park Joy*. Bottom row from left to right: *Pedestrian Area* and *RaceHorses*.

Macroblocks with saliency above the 75-th percentile threshold were considered as ROIs. Hence, in our experiments, ROI occupies up to 25% of the frame.

To simulate a real video streaming scenario with RECAP as its error control mechanism, a video frame was selected randomly, and its macroblocks in non-ROI parts were dropped randomly based on a two-state Gilbert model [34] at two different average loss rates (3% and 10%) with an average burst loss length of 8. The corrupted frame was then concealed using a conventional direct frame copy method (FC), the original RECAP algorithm, our previous error concealment method from [12], as well as our proposed error concealment method from Section III-A. The direct frame copy method conceals a lost block by copying the co-located block from the previous frame. To generate the results, it was assumed that the last correctly-received reference frame was $d = 5$ or $d = 10$ frames away. In practice, this distance is random. We used $d = 5$ and $d = 10$ simply as representative test values.

1) Objective Quality Assessment: In Table III, we compared the performance of the proposed error concealment method with the direct frame copy method (FC), the RECAP method and our earlier error concealment method from [12] at 3% loss rate, and at two reference frame distances, $d = 5$ and $d = 10$, based on three metrics: PSNR, SSIM [35], and VQM [36], [37]. These frame-level metrics were computed at the aforementioned average loss rates on the concealed frames. The general VQM model [36] was utilized for computing the VQM values. Only the luma channel was considered for computing the PSNR and SSIM values. Table IV shows the results for the same comparison as in Table III but at an average loss rate of 10%.

As seen from Tables III and IV, the proposed method is able to improve the PSNR of the concealed frames by several dBs compared to FC, by up to 3.6 dB compared to RECAP (*Crew* with $d = 10$ at 10% loss rate), and by up to 0.7 dB compared to our earlier method from [12] (*Tractor* with $d = 5$ at 3% loss rate). This shows that the proposed error concealment method is able to provide correct side information for resolving the ambiguity in reconstructing the missing blocks in the under-determined problem of error concealment. As seen from the SSIM and VQM results, the proposed error concealment method provides better quality than either RECAP or FC. Note that the smaller the VQM value, the better the quality. We also note that the objective quality of the proposed error concealment method as measured by the SSIM and VQM metrics is close to or better than our earlier method from [12].

TABLE III
COMPARING THE PROPOSED ERROR CONCEALMENT METHOD WITH FC, RECAP, AND [12], USING VARIOUS IMAGE/VIDEO QUALITY ASSESSMENT METHODS, AT 3% LOSS RATE, AND AT TWO REFERENCE FRAME DISTANCES: $d = 10$ AND $d = 5$

Sequence	Method	$d = 5$			$d = 10$		
		PSNR	SSIM	VQM	PSNR	SSIM	VQM
Soccer	FC	25.4	0.896497	0.293116	24.1	0.885284	0.310229
	RECAP	35.7	0.955522	0.168544	34.6	0.943254	0.204069
	[12]	36.1	0.956073	0.153910	35.3	0.944754	0.161319
	Proposed	36.2	0.956709	0.139484	35.5	0.946126	0.157103
RaceHorses	FC	31.5	0.956102	0.187969	30.4	0.951865	0.210391
	RECAP	38.3	0.990184	0.077677	37.6	0.978554	0.088587
	[12]	38.5	0.992794	0.077879	38.0	0.979591	0.081535
	Proposed	38.9	0.993050	0.077382	38.2	0.980300	0.080902
Tractor	FC	27.8	0.889505	0.260994	26.3	0.877193	0.318149
	RECAP	34.2	0.973539	0.125716	32.4	0.930999	0.170746
	[12]	34.3	0.981457	0.126531	33.8	0.940424	0.139316
	Proposed	35.0	0.982164	0.125092	34.2	0.943242	0.138356
Crew	FC	30.3	0.924047	0.288189	28.3	0.912229	0.297958
	RECAP	36.5	0.965405	0.185925	35.1	0.956591	0.217980
	[12]	38.6	0.972251	0.137344	38.2	0.969242	0.148807
	Proposed	38.9	0.973920	0.136188	38.4	0.970139	0.148376
Park Joy	FC	22.5	0.810948	0.347995	21.3	0.787842	0.349853
	RECAP	28.3	0.911193	0.191236	27.1	0.890651	0.228965
	[12]	28.5	0.915182	0.187241	27.6	0.892059	0.206739
	Proposed	28.8	0.919213	0.182359	27.8	0.892895	0.201442
Pedestrian Area	FC	26.2	0.912223	0.409675	24.5	0.889103	0.445602
	RECAP	34.7	0.955491	0.336493	33.7	0.945028	0.371565
	[12]	37.5	0.966223	0.289672	37.0	0.961154	0.308278
	Proposed	37.8	0.969751	0.223417	37.2	0.961974	0.305239

TABLE IV
COMPARING THE PROPOSED ERROR CONCEALMENT METHOD WITH FC, RECAP, AND [12], USING VARIOUS IMAGE/VIDEO QUALITY ASSESSMENT METHODS, AT 10% LOSS RATE, AND AT TWO REFERENCE FRAME DISTANCES: $d = 10$ AND $d = 5$

Sequence	Method	$d = 5$			$d = 10$		
		PSNR	SSIM	VQM	PSNR	SSIM	VQM
Soccer	FC	20.3	0.743474	0.357723	19.3	0.723996	0.355943
	RECAP	31.1	0.874573	0.366916	30.1	0.847758	0.413155
	[12]	31.5	0.875573	0.315949	31.0	0.854268	0.339648
	Proposed	31.6	0.876853	0.305230	31.2	0.855404	0.313936
RaceHorses	FC	22.5	0.761886	0.434774	21.2	0.738031	0.432532
	RECAP	29.1	0.878562	0.276700	28.5	0.857703	0.326603
	[12]	29.5	0.883668	0.238212	28.9	0.863775	0.264199
	Proposed	29.7	0.885283	0.235563	29.1	0.866455	0.240614
Tractor	FC	23.0	0.694299	0.357006	21.5	0.661935	0.428525
	RECAP	30.0	0.905805	0.248732	27.2	0.790720	0.271512
	[12]	30.4	0.906624	0.220100	28.9	0.822753	0.225952
	Proposed	30.5	0.910247	0.203267	29.0	0.822901	0.223710
Crew	FC	24.4	0.790344	0.401263	23.2	0.768557	0.412373
	RECAP	31.1	0.890988	0.320792	29.8	0.867791	0.363882
	[12]	33.3	0.908552	0.261871	33.2	0.904089	0.292193
	Proposed	33.6	0.910276	0.241848	33.4	0.904329	0.253789
Park Joy	FC	20.1	0.675721	0.481654	18.5	0.635485	0.463155
	RECAP	24.3	0.832451	0.212345	23.2	0.813476	0.240211
	[12]	24.7	0.840101	0.202015	23.7	0.820121	0.233451
	Proposed	24.9	0.843526	0.198722	24.0	0.823456	0.229312
Pedestrian Area	FC	24.0	0.876393	0.588472	22.1	0.841600	0.594122
	RECAP	32.5	0.918282	0.211396	31.6	0.912373	0.243142
	[12]	35.8	0.941829	0.172335	34.7	0.932336	0.185134
	Proposed	36.0	0.942882	0.170112	35.1	0.939124	0.178952

The results demonstrate that even though our earlier method in [12] used the actual IKN saliency while the method proposed here uses only an approximation, we are able to improve upon the results from [12]. This is because the present error concealment formulation in (1) allows for direct search for the missing block $\mathbf{X}$, whereas in [12], the concealment proceeded indirectly by applying saliency reduction operators, in an iterative fashion, upon RECAP candidate blocks. This, combined with the non-convexity of the objective function from [12], made the algorithm in [12] susceptible to getting stuck in a local optimum. The present algorithm does not have that problem, and it is computationally more efficient.

2) Subjective Evaluation: Since our proposed error concealment method aims at reducing the saliency of concealed regions, we performed a subjective test to verify the improvement in subjective quality. For this purpose, we compared the subjective quality of our proposed error concealment method with the RECAP method as well as our earlier method proposed in [12].

In our experiment, a Two Alternative Forced Choice (2AFC) method [38] was used to compare subjective video quality. In 2AFC, the participant is asked to make a choice between two alternatives, in this case the original RECAP method or our previous method from [12], and the method proposed in this paper. This way of comparing image 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) [39]. The 2AFC method can be implemented either spatially (i.e., showing the two alternatives side by side at the same time period) or temporally (i.e., presenting one alternative after another in two consecutive time intervals). In our experiments, we used the latter approach for *Park Joy* and *Pedestrian Area* due to their high resolution while we used the former approach for the other four test sequences.

For obtaining the subjective results for *Soccer*, *Race Horses*, *Tractor*, and *Crew*, in each trial, participants were looking at two side-by-side videos (in the same vertical position, and separated by 1 cm horizontally) on a mid-gray background. Each video pair was shown for 9 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). For obtaining the subjective results for *Park Joy* and *Pedestrian Area*, in each trial, two videos were presented to the participants one after the other, with a one-second mid-gray background in between. After the second video, a mid-gray background was displayed for 5 seconds, and the participants were asked to indicate on the answer sheet, which of the two videos look better (First or Second). In all the trials, the participants were asked to answer either Left (First) or Right (Second) for each video pair, regardless of how certain they were of their response. Participants did not know which video was obtained by our proposed method and which one was obtained by the alternative method (RECAP or [12]). Randomly chosen half of the trials had the video produced by our method on the left side of the screen (or in the first time interval in the case of high-resolution sequences) and the other half on the right side (in the second time interval). This gave a total of $6 \times 2 \times 2 = 24$ trials for comparing the proposed method against RECAP, and another 24 trials for comparing the proposed method against our earlier method from [12].

TABLE V
COMPARING THE PROPOSED ERROR CONCEALMENT METHOD WITH THE RECAP METHOD BASED ON THE SUBJECTIVE RESULTS AT 2 DIFFERENT AVERAGE LOSS RATES

Loss Rate	Method	*Crew*	*Soccer*	*Tractor*	*Race Horses*	*Park Joy*	*Pedestrian Area*
3%	RECAP	2	6	6	3	8	2
	Proposed Method	28	24	24	27	22	28
	p-value	0.0001	0.0010	0.0010	0.0001	0.0106	0.0001
10%	RECAP	1	5	8	4	6	4
	Proposed Method	29	25	22	26	24	26
	p-value	0.0001	0.0003	0.0106	0.0001	0.0010	0.0001

TABLE VI
COMPARING THE PROPOSED ERROR CONCEALMENT METHOD WITH OUR EARLIER METHOD IN [12] BASED ON THE SUBJECTIVE RESULTS AT 2 DIFFERENT AVERAGE LOSS RATES

Loss Rate	Method	*Crew*	*Soccer*	*Tractor*	*Race Horses*	*Park Joy*	*Pedestrian Area*
3%	[12]	11	9	10	12	13	10
	Proposed Method	19	21	20	18	17	20
	p-value	0.1441	0.0285	0.0679	0.2733	0.4652	0.0679
10%	[12]	17	10	16	14	11	12
	Proposed Method	13	20	14	16	19	18
	p-value	0.4652	0.0679	0.7150	0.7150	0.1441	0.2733

The experiment was run in a quiet room with 15 participants (all male except four, and of age between 18 and 30). All participants had normal or corrected to normal vision. A 26-inch Dell monitor with brightness 300 cd/m^2 and resolution 1920×1080 pixels was used in our experiments. The brightness and contrast of the monitor were set to 75%. 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 10 minutes.

The results for the comparison between the RECAP method and our proposed error concealment method are shown in Table V, where we indicate the number of responses that showed preference for the original RECAP method and the proposed method at the two tested average loss rates. Similarly, the results for the comparison between our earlier method in [12] and our proposed error concealment method in this paper are shown in Table VI, where we indicate the number of responses that showed preference for the method in [12] and the proposed method at the two tested average loss rates.

We used the two-sided chi-squared (χ^2) test [31] to examine the statistical significance of the results. In Table V, the null hypothesis is that there is no preference for either the RECAP method or the proposed method. Similarly, the null hypothesis in Table VI is that there is no preference for either the method in [12] or the proposed method in this paper. Under this hypothesis, the expected number of votes is 15 for each method under study. The p-value [31] is also indicated in these two tables. In experimental sciences, as a rule of thumb, the null hypothesis is rejected when $p < 0.05$. When this happens in Table V or 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.

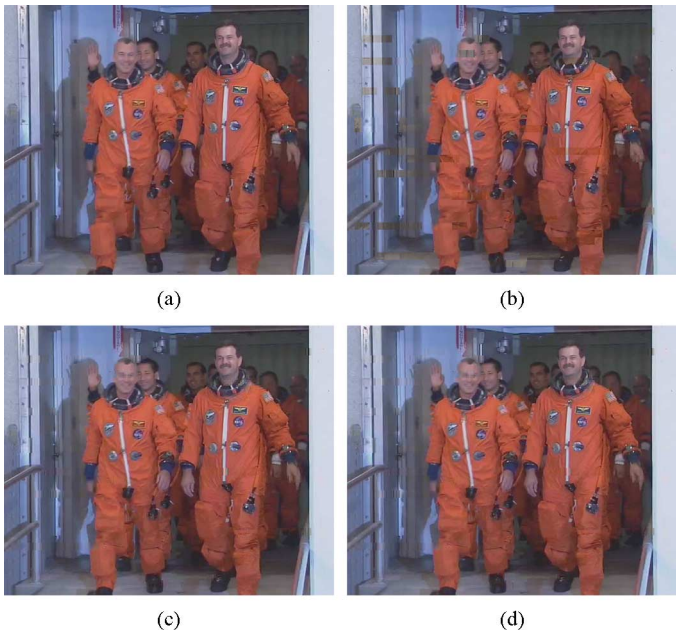


Fig. 7. Comparing the visual quality of the proposed error concealment method with the RECAP method and the method proposed in [12] on *Crew*. (a) original video frame, (b) the reconstructed frame using the RECAP method based on a reference frame, which is 10 frames away from the current frame ($\mathrm{PSNR} = 34.34\ \mathrm{dB}$), (c) the reconstructed frame using [12] ($\mathrm{PSNR} = 36.55\ \mathrm{dB}$), (d) the reconstructed frame using the proposed error concealment method ($\mathrm{PSNR} = 36.83\ \mathrm{dB}$).

As seen in Table V, in all of the 24 trials the p-value is smaller than 0.05, which indicates that subjects showed a statistically significant preference for our proposed method over RECAP. Looking across all trials (i.e., summing up all the votes for the two options), the results show that participants have preferred our method more than the RECAP method (305 vs. 55 votes) with overall $p = 0.0001$, which is a very statistically significant result, because the odds of it occurring by chance are 1 in 10000. This confirms that the proposed method is able to improve the perceptual quality of the concealed frames compared to the original RECAP method. As seen in Table VI, in all of the 24 trials except for one (*Soccer* at 3%) the p-value is above 0.05, which indicates that the subjective quality of our proposed error concealment method is similar to our earlier method in [12] on all videos except for *Soccer* at 3% loss rate, in which case our new method provided statistically better subjective quality.

Fig. 7 shows an illustration of the visual quality of frames concealed by the method proposed in this paper compared to those produced by the original RECAP method and those produced by our earlier method from [12] on *Crew*. One can easily see that our new method is able to improve the visual quality of the concealed frames compared to the original RECAP method. We also note that the visual quality of the frame produced by the new method is similar to that produced by our earlier method from [12].

VI. Conclusion

Error concealment in loss-corrupted streaming video is a challenging under-determined problem. In this paper, we add a low-saliency prior as a regularization term to the replacement block search problem. Low saliency provides the side information in ROI-based UEP video streaming systems for client to identify correct replacement blocks for concealment. Also, low saliency reduces viewer's visual attention on the loss-stricken regions. Incorporated into a previously proposed RECAP error concealment setup, our experimental results show that our method can clearly improve the visual quality of the loss-corrupted frames both objectively (up to 3.6 dB in PSNR) and subjectively. Moreover, incorporating the newly-developed convex approximation to visual saliency into the error concealment process results in significant complexity reduction, while at the same time providing a gain of up to 0.7 dB in PSNR compared to an earlier version of the algorithm.

Appendix

In Section III-B.2, we introduced a matrix expansion operator $\mathcal{Z}(\mathbf{X}, \mathcal{N}(\mathbf{X}))$ for expanding a $m \times m$ matrix $\mathbf{X}$ to a $p \times p$ matrix $\mathbf{X}_e$ by zero-padding it based on the $p \times p$ spatial neighborhood $\mathcal{N}(\mathbf{X})$. As we mentioned in Section III-B.2, $\mathcal{Z}(\mathbf{X}, \mathcal{N}(\mathbf{X}))$ can be realized by a linear transformation as follows:

$$\mathbf{X}_e = \mathcal{Z}(\mathbf{X}, \mathcal{N}(\mathbf{X})) = \mathbf{M}\mathbf{X}\mathbf{N}, \tag{17}$$

where $\mathbf{M}$ is a binary matrix of size $p \times m$ and $\mathbf{N}$ is a binary matrix of size $m \times p$. In this appendix, we derive $\mathbf{M}$ and $\mathbf{N}$ so that $\mathcal{Z}(\mathbf{X}, \mathcal{N}(\mathbf{X}))$ can be utilized in our error concealment methodology proposed in Section III-A.

Note that in our proposed error concealment method, missing blocks are reconstructed in a raster-scan order. Hence, the causal neighbors of all missing blocks will always be available. However, the anti-causal neighbors of the missing blocks may be missing. Therefore, depending on the availability of the anti-causal neighbors of the missing block in a 8-connected neighborhood, we may encounter one of the cases depicted in Fig. 8. For each of these cases, we obtain a different $\mathbf{M}$ and $\mathbf{N}$ as follows:

- Case 1 in Fig. 8 shows the situation in which all the 8-connected neighbors of the current block are available, and we want to expand the current block $\mathbf{X}$ by zero-padding it from all sides. In this case, $\mathbf{M}$ and $\mathbf{N}$ are defined as follows:

$$\mathbf{M} = \begin{pmatrix} [\mathbf{0}]_{(p-N_b)/2 \times N_b} \\ [\mathbf{I}]_{N_b \times N_b} \\ [\mathbf{0}]_{(p-N_b)/2 \times N_b} \end{pmatrix}_{p \times N_b}, \quad \mathbf{N} = \mathbf{M}^t, \tag{18}$$

  where $[\mathbf{0}]_{x \times y}$ denotes a $x \times y$ matrix whose elements are all zero, and $[\mathbf{I}]_{x \times y}$ denotes the identity matrix of size $x \times y$.
- Case 2 in Fig. 8 shows the situation in which only the causal 8-connected neighbors of the current block $\mathbf{X}$ are available, and we want to expand the current block by zero-padding it from the top and left. In this case, $\mathbf{M}$ and $\mathbf{N}$ are defined as follows:

$$\mathbf{M} = \begin{pmatrix} [\mathbf{0}]_{(p-N_b) \times N_b} \\ [\mathbf{I}]_{N_b \times N_b} \end{pmatrix}_{p \times N_b}, \mathbf{N} = \mathbf{M}^t. \tag{19}$$

- Case 3 in Fig. 8 shows the situation in which all the 8-connected neighbors of the current block $\mathbf{X}$ are available except for one or more of its neighbors from below, and we

want to expand the current block by zero-padding it from the left, top, and right. In this case, $\mathbf{M}$ and $\mathbf{N}$ are defined as follows:

$$\mathbf{M} = \begin{pmatrix} [\mathbf{0}]_{(p-N_b)\times N_b} \\ [\mathbf{I}]_{N_b\times N_b} \end{pmatrix}_{p\times N_b}, \tag{20}$$

$$\mathbf{N} = \begin{pmatrix} [\mathbf{0}]_{N_b\times(p-N_b)/2} & [\mathbf{I}]_{N_b\times N_b} & [\mathbf{0}]_{N_b\times(p-N_b)/2} \end{pmatrix}_{N_b\times p}. \tag{21}$$

- Case 4 in Fig. 8 shows the situation in which all 8-connected neighbors of the current block are available except for one or more of its 8-connected neighbors to the right, and we want to expand the current block $\mathbf{X}$ by zero-padding it from the left, top, and bottom. In this case, $\mathbf{M}$ and $\mathbf{N}$ are defined as follows:

$$\mathbf{M} = \begin{pmatrix} [\mathbf{0}]_{(p-N_b)/2\times N_b} \\ [\mathbf{I}]_{N_b\times N_b} \\ [\mathbf{0}]_{(p-N_b)/2\times N_b} \end{pmatrix}_{p\times N_b}, \tag{22}$$

$$\mathbf{N} = \begin{pmatrix} [\mathbf{0}]_{N_b\times(p-N_b)} & [\mathbf{I}]_{N_b\times N_b} \end{pmatrix}_{N_b\times p}. \tag{23}$$

- Case 5 in Fig. 8 shows the situation in which the current block is on the left boundary of the frame and all of its 8-connected neighbors are available, and we want to expand the current block by zero-padding it from all sides except left. In this case, $\mathbf{M}$ and $\mathbf{N}$ are defined as follows:

$$\mathbf{M} = \begin{pmatrix} [\mathbf{0}]_{(p-N_b)/2\times N_b} \\ [\mathbf{I}]_{N_b\times N_b} \\ [\mathbf{0}]_{(p-N_b)/2\times N_b} \end{pmatrix}_{p\times N_b}, \tag{24}$$

$$\mathbf{N} = \begin{pmatrix} [\mathbf{I}]_{N_b\times N_b} & [\mathbf{0}]_{N_b\times(p-N_b)} \end{pmatrix}_{N_b\times p}. \tag{25}$$

- Case 6 in Fig. 8 shows the situation in which the current block is at the top-right corner of the frame and all of its 8-connected neighbors are available, and we want to expand the current block by zero-padding it from the left and bottom. In this case, $\mathbf{M}$ and $\mathbf{N}$ are defined as follows:

$$\mathbf{M} = \begin{pmatrix} [\mathbf{I}]_{N_b\times N_b} \\ [\mathbf{0}]_{(p-N_b)\times N_b} \end{pmatrix}_{p\times N_b}, \tag{26}$$

$$\mathbf{N} = \begin{pmatrix} [\mathbf{0}]_{N_b\times(p-N_b)} & [\mathbf{I}]_{N_b\times N_b} \end{pmatrix}_{N_b\times p}. \tag{27}$$

- Case 7 in Fig. 8 shows the situation in which the current block is at the top-left corner of the frame and all of its 8-connected neighbors are available, and we want to expand the current block by zero-padding it from the right and bottom. In this case, $\mathbf{M}$ and $\mathbf{N}$ are defined as follows:

$$\mathbf{M} = \begin{pmatrix} [\mathbf{I}]_{N_b\times N_b} \\ [\mathbf{0}]_{(p-N_b)\times N_b} \end{pmatrix}_{p\times N_b}, \tag{28}$$

$$\mathbf{N} = \begin{pmatrix} [\mathbf{I}]_{N_b\times N_b} & [\mathbf{0}]_{N_b\times(p-N_b)} \end{pmatrix}_{N_b\times p}. \tag{29}$$

- Case 8 in Fig. 8 shows the situation in which the current block is at the bottom-left corner of the frame and all of its 8-connected neighbors are available, and we want to expand the current block by zero-padding it from the top and right. In this case, $\mathbf{M}$ and $\mathbf{N}$ are defined as follows:

$$\mathbf{M} = \begin{pmatrix} [\mathbf{0}]_{(p-N_b)\times N_b} \\ [\mathbf{I}]_{N_b\times N_b} \end{pmatrix}_{p\times N_b}, \tag{30}$$

$$\mathbf{N} = \begin{pmatrix} [\mathbf{I}]_{N_b\times N_b} & [\mathbf{0}]_{N_b\times(p-N_b)} \end{pmatrix}_{N_b\times p}. \tag{31}$$

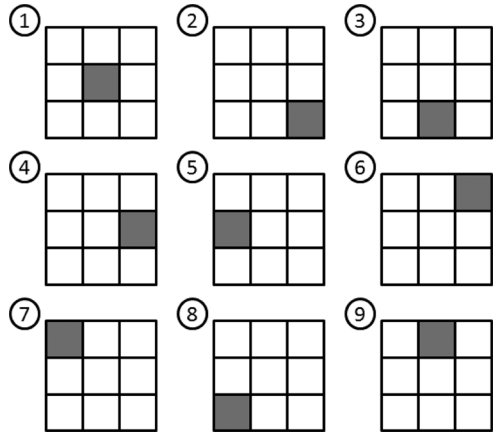


Fig. 8. In the proposed error concealment method, depending on the availability of the neighbors of a missing block, various situations may arise. In this figure, the missing block has been depicted by a gray box while its available neighbors have been depicted by white boxes. The available neighbors of the missing block $\mathbf{X}$ are used to define $\mathcal{N}(\mathbf{X})$.

- Case 9 in Fig. 8 shows the situation in which the current block in on the top boundary of the frame and all of its 8-connected neighbors are available, and we want to expand the current block by zero-padding it from all sides except the top. In this case, $\mathbf{M}$ and $\mathbf{N}$ are defined as follows:

$$\mathbf{M} = \begin{pmatrix} [\mathbf{I}]_{N_b\times N_b} \\ [\mathbf{0}]_{(p-N_b)\times N_b} \end{pmatrix}_{p\times N_b}, \tag{32}$$

$$\mathbf{N} = \begin{pmatrix} [\mathbf{0}]_{N_b\times(p-N_b)/2} & [\mathbf{I}]_{N_b\times N_b} & [\mathbf{0}]_{N_b\times(p-N_b)/2} \end{pmatrix}_{N_b\times p}. \tag{33}$$

For all other possible cases, we assume that $\mathcal{N}(\mathbf{X})$ covers only $\mathbf{X}$, and so we set both $\mathbf{M}$ and $\mathbf{N}$ to $N_b \times N_b$ identity matrices.

REFERENCES

[1] Cisco Visual Networking Index: Forecast and Methodology 2010–2015. [Online]. Available: http://www.cisco.com/.

[2] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multi-view imaging and 3DTV," *IEEE Signal Process. Mag.*, vol. 24, no. 6, pp. 10–21, Nov. 2007.

[3] G.-M. Muntean, G. Ghinea, and T. Sheehan, "Region of interest-based adaptive multimedia streaming scheme," *IEEE Trans. Broadcast.*, vol. 54, no. 2, pp. 296–303, Jun. 2008.

[4] F. Boulos, W. Chen, B. Parrein, and P. L. Callet, "A new H.264/AVC error resilience model based on regions of interest," in *Proc. 17th Int. Packet Video Workshop, PV 2009*, Seattle, WA, USA, May 2009.

[5] N. Bruce and P. Kornprobst, "Region-of-interest intra prediction for H.264/AVC error resilience," in *Proc. IEEE Int. Conf. Image Process.*, Cairo, Egypt, Nov. 2009.

[6] 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.

[7] 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.

[8] Oxford Dictionary: Definition of Concealment. [Online]. Available: http://oxforddictionaries.com/definition/american_english/conceal.

[9] Y. Chen, Y. Hu, O. Au, H. Li, and C. W. Chen, "Video error concealment using spatio-temporal boundary matching and partial differential equation," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 2–15, Jan. 2008.

[10] C. Yeo, W.-T. Tan, and D. Mukherjee, "Receiver error concealment using acknowledge preview (RECAP)–An approach to resilient video streaming," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, Taipei, Taiwan, Apr. 2009.

[11] 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.

[12] H. Hadizadeh, I. V. Bajić, and G. Cheung, "Saliency-cognizant error concealment in loss-corrupted streaming video," in *Proc. 2012 IEEE Int. Conf. Multimedia & Expo (ICME 2012)*, Jul. 2012, pp. 73–79.

[13] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, Minneapolis, MN, USA, Jun. 2007.

[14] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, Miami Beach, FL, USA, Jun. 2009.

[15] 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'12 Workshop: QoEMC*, Anaheim, CA, USA, Dec. 2012, pp. 1304–1308.

[16] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ, USA: Prentice Hall, 1993.

[17] H. Hadizadeh, I. V. Bajić, and G. Cheung, Complexity of Saliency-Cognizant Error Concealment based on the Itti-Koch-Niebur Saliency Model, Multimedia Communications Lab, Simon Fraser Univ., 2012. [Online]. Available: http://summit.sfu.ca/item/10942.

[18] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*. Philadelphia, PA, USA: SIAM Studies in Applied Mathematics, 1994.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[20] J. W. Woods, *Multidimensional Signal, Image, and Video Processing and Coding*, 2nd ed. New York, NY, USA: Academic/Elsevier, 2012.

[21] H. Stark and J. W. Woods, *Probability, Statistics, and Random Processes for Engineers*, 4th ed. Upper Saddle River, NJ, USA: Pearson/Prentice Hall, 2012.

[22] J. S. Lim, *Two-Dimensional Signal and Image Processing*. Englewood Cliffs, NJ, USA: Prentice Hall, 1989.

[23] A. Torralba and A. Oliva, "Statistics of natural image categories," *Netw.: Comput. Neural Syst.*, vol. 14, pp. 391–412, 2003.

[24] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York, NY, USA: Wiley, 1999.

[25] R. W. Vuduc, "Automatic performance tuning of sparse matrix kernels," Ph.D. dissertation, Univ. California, Computer Science, Berkeley, CA, USA, 2003.

[26] N. D. B. Bruce and J. K. Tsotsos, "Saliency based on information maximization," *Adv. Neural Inf. Process. Syst.*, pp. 155–162, Jun. 2006.

[27] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2009.

[28] Saliency Benchmark Datasets. [Online]. Available: http://people.csail.mit.edu/tjudd/SaliencyBenchmark/.

[29] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *Vision Res.*, vol. 9(12), no. 10, pp. 1–15, 2009.

[30] B. Tatler, R. Baddeley, and I. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vision R.*, vol. 45, no. 5, pp. 643–659, 2005.

[31] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*. London, U.K.: Chapman & Hall/CRC, 2007.

[32] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.

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

[34] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Syst. Tech. J.*, vol. 39, pp. 1253–1266, Sep. 1960.

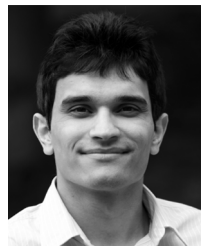
[35] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[36] 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.

[37] 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.

[38] M. Taylor and C. Creelman, "PEST: Efficient estimates on probability functions," *J. Acoust. Soc. Amer.*, vol. 41, pp. 782–787, 1967.

[39] ITU-R, "Recommendation BT.500-8: Methodology for the Subjective Assessment of the Quality of Television Pictures," Tech. Rep., ITU, 1998.



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

His research interests include perceptual image/video coding, visual attention modelling, error resilient video transmission, image/video processing, computer vision, data mining, and machine learning. He was the recipient of the Best Paper Runner-up Award at ICME 2012 in Melbourne, Australia, and the Microsoft Research and Canon Information Systems Research Australia (CiSRA) Student Travel Grant for ICME 2012. He is currently a Research Associate at Simon Fraser University, 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 Rensselaer Polytechnic Institute, Troy, NY, in 2000, 2002, and 2003, respectively.

He is currently Associate Professor of Engineering Science at Simon Fraser University, Burnaby, BC, Canada. His research interests include signal, image, and video processing and compression, perceptual aspects of information processing, and multimedia communications. He has authored about 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 the 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.

Gene Cheung (M'00–SM'07) received the B.S. degree in electrical engineering from Cornell University in 1995, and the M.S. and Ph.D. degrees in electrical engineering and computer science from the University of California, Berkeley, in 1998 and 2000, respectively.

He was a senior researcher in Hewlett-Packard Laboratories Japan, Tokyo, from 2000 until 2009. He is now an associate professor in National Institute of Informatics in Tokyo, Japan.

His research interests include 3D visual representation, single-/multiple-view video coding and streaming, and immersive communication. He has published over 100 international conference and journal publications. He has served as associate editor for IEEE Transactions on Multimedia from 2007 to 2011 and currently serves as associate editor for DSP Applications Column in IEEE Signal Processing Magazine and APSIPA journal on signal and information processing, and as area editor for EURASIP Signal Processing: Image Communication. He currently serves as member of the Multimedia Signal Processing Technical Committee (MMSP-TC) in IEEE Signal Processing Society (2012–2014). He has also served as area chair in IEEE International Conference on Image Processing (ICIP) 2010, 2012–2013, technical program co-chair of International Packet Video Workshop (PV) 2010, track co-chair for Multimedia Signal Processing track in IEEE International Conference on Multimedia and Expo (ICME) 2011, symposium co-chair for CSSMA Symposium in IEEE GLOBECOM 2012, and area chair for ICME 2013. He is invited as plenary speaker for IEEE International Workshop on Multimedia Signal Processing (MMSP) 2013 on the topic "3D visual communication: media representation, transport and rendering". He is a co-author of best student paper award in IEEE Workshop on Streaming and Media Communications 2011 (in conjunction with ICME 2011), best paper finalists in ICME 2011 and ICIP 2011, and best paper runner-up award in ICME 2012.