SALIENCY-COGNIZANT ERROR CONCEALMENT IN LOSS-CORRUPTED STREAMING VIDEO

Hadi Hadizadeh†, Ivan V. Bajić†, and Gene Cheung‡

†Simon Fraser University, Burnaby, BC, Canada, ‡National Institute of Informatics, Tokyo, Japan

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 right 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 [1] 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 design and apply four saliency reduction operators iteratively in a loop, in order to reduce the saliency of candidate blocks. Experimental results show that: i) PSNR of the error-concealed frames can be increased dramatically (up to 3.2dB 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*—** Video streaming, error concealment, visual saliency

1. 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 [2], 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 the timing constraint of streaming video (i.e., a video packet arriving at decoder past its playback deadline is useless). Third, new media types such as ultra-high-resolution video and multiple-view video [3] that promise enhancement of viewing experience are also further straining resource-limited networks due to their large sizes. Under these practical constraints, it is very difficult to guarantee error-free delivery of the entire video from sender to receiver in a timely manner.

Many previous works [4, 5, 6] employed the pro-active methodology of unequal error protection (UEP) of video data, where important packets are protected more heavily (e.g., using stronger Forward Error Correction (FEC) codes). Typically, more important packets contain viewer's probable Regions-of-Interest (ROI) [7] in a video frame, or regions with higher *visual saliency* [8]—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. In this paper, we study the complementary problem of *error concealment*: given the occasional unavoidable 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.

Error concealment is typically an under-determined problem: there are insufficient number of well-defined criteria (e.g., 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 overall subjective quality.

We study the effectiveness of a low-saliency prior in the context of a previously proposed RECAP error concealment system [1]. 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 design and apply four saliency reduction operators iteratively in a loop, leveraging on previous work on saliency manipulation such as [10], so that the saliency of candidate blocks is reduced. Experimental results show that: i) PSNR of the error-concealed frames can be increased dramatically (up to 3.2dB 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.

This work was supported in part by the NSERC grant RGPIN 327249.

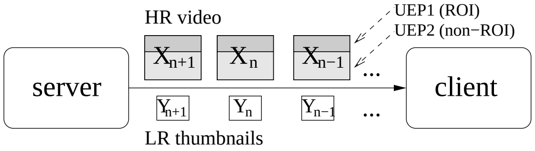


Fig. 1. Overview of RECAP packet loss recovery system.

The outline of the paper is as follows. We first discuss related work in Section 2. We then present an overview of the RECAP video transmission system and our chosen visual saliency model in Sections 3 and 4, respectively. We discuss our proposed error concealment strategy with low-saliency prior in Section 5. Finally, experimental results and conclusions are presented in Sections 6 and 7, respectively.

2. RELATED WORK

In the face of challenging network conditions during real-time video streaming, UEP strategies [4, 5, 6] protect visually important (salient) regions more heavily. An often overlooked question in these works is how to conceal missing blocks in the less important regions when packet losses do occur? If concealment is done in a *saliency-myopic* way, so that the resulting salient features draw attention to the (likely) imperfectly recovered 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 can be done in a *saliency-cognizant* manner, resulting in recovered blocks that do not draw unnecessary attention.

We note that though we apply our low-saliency prior to the RECAP video transmission system [1] (where LR thumbnails are transmitted from server to client to facilitate loss recovery) in this paper for concreteness, we believe 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 spatial regions because that is the *typical* case. Packet losses in more heavily protected high-saliency spatial regions, while possible, is a *rare* case, and hence will not affect much the average performance of the system as long as some default concealment scheme is performed.

Visual saliency—a measure of propensity for drawing visual attention—has been a subject of intense study in the past decade [8, 11, 12]. While it is debatable which proposed method computes the most accurate saliency maps at reasonable complexity, we selected [8], [13] for our ground truth saliency calculation due to its wide acceptability and potential parallelism in implementation for our video streaming scenario. Note that the focus of our paper is not on new computational models of saliency, but on the application of saliency analysis to error concealment. While earlier works have applied visual saliency principles for video compression [13], to the best of our knowledge, we are the first to apply saliency analysis for error concealment of streaming video.

3. RECAP VIDEO TRANSMISSION SYSTEM

We first present an overview of the RECAP video transmission system [1], 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 details in Section 6). 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.

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 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, 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 [1] 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. First, we discuss the visual saliency model we selected.

4. OVERVIEW OF THE VISUAL SALIENCY MODEL

Among the existing bottom-up computational models of visual attention, the Itti-Koch-Niebur (IKN) model [8] 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 subsample the input image, yielding an image-size-reduction factor ranging from 1:1 (scale zero) to 1:256 (scale eight) in eight octaves [8].

The contrast in each feature channel is then computed using a "center-surround" mechanism, which is implemented in the model as the difference between fine and coarse scales: the center is a pixel at scale $c \in \{2, 3, 4\}$, and the surround is the corresponding pixel at scale $s = c + d$, with $d \in \{3, 4\}$. The across-scale difference between two levels of the pyramid is obtained by interpolation to the finer scale and point-by-point subtraction. The obtained contrast (feature) maps are then combined across scales through a 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. An extra motion and flicker channel can also be added to the IKN model in order to make it more suitable for video [13].

Note that 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 of the image. This yields the normalized frequency spectrum of the image at level 8 to be in range $[0 - \pi/256]$. Also, since the conspicuity maps are resized to level 4 before combination, this results in using the frequency content of the original image in the range $[\pi/256 - \pi/16]$ by the IKN model [12]. We will use this fact in one part of our proposed method.

5. LOW-SALIENCY PRIOR IN ERROR CONCEALMENT

Having reviewed the RECAP video transmission system and our chosen saliency model in Sections 3 and 4 respectively, we now discuss how we incorporate a low-saliency prior into the RECAP error concealment scheme in mathematical details.

Let y be the thumbnail of a lost slice x, i.e., the low-pass-filtered and down-sampled version of x. Thumbnail y is subsequently compressed to $\tilde{y}$ before transmission. $\tilde{y}$ is hence related to x as follows:

$$\tilde{y} = DLx + \epsilon, \tag{1}$$

where L is a low-pass filter to avoid aliasing before down-sampling, D is the down-sample operator, and ϵ is quantization noise in $\tilde{y}$ due to lossy compression. We assume $\tilde{y}$ is correctly decoded, and the goal is to use low-pass information in $\tilde{y}$ to recover the lost slice x. In general, this is an under-determined system of equations, and there are many candidates $\hat{x}$'s that yield small l_2-norm $\|\tilde{y} - DL\hat{x}\|_2$.

To resolve this ambiguity, we introduce an additional low-saliency regularization term $\lambda S(\hat{x})$ to the l_2-norm as objective, where $S(\hat{x})$ is the the visual saliency of reconstructed slice $\hat{x}$ in the video frame:

$$\min_{\hat{x}} \ \|\tilde{y} - DL\hat{x}\|_2 + \lambda S(\hat{x}), \tag{2}$$

and λ is a non-negative weight parameter that trades off the relative importance of the l_2-norm and visual saliency. In the remainder of this section, we discuss how (2) can be solved efficiently.

5.1. Algorithm Overview

We first present our strategy to find a good replacement slice $\hat{x}$ in (2) for a loss-corrupted frame. We divide the missing slice $\hat{x}$ into smaller 16×16 MBs $\hat{x}_i$'s, such that $\bigcup_i \hat{x}_i = \hat{x}$. For each MB $\hat{x}_i$, the best M candidate blocks $\hat{x}_i^{(m)}$ ($1 \leq m \leq M$) from a HR correctly-received reference frame that have the smallest l_2-norm error with respect to the corresponding thumbnail block of the current MB (i.e., $\tilde{y}_i$) are first identified. We also include the 4 adjacent spatial neighbors in the causal neighborhood of the current MB in the search process for finding the best M candidate blocks. The causal neighbors may be previously concealed. The search procedure used here is the same as the search procedure used in the original RECAP algorithm as proposed in [1]. All the chosen M candidate blocks are then examined based on the objective function (2) to select only $K < M$ of them as the final candidate blocks. Note that examining the objective function (2) on all possible candidate blocks in a search region inside the HR reference frame might be very time consuming due to the saliency computation step. However, the above approach can reduce the computational complexity significantly.

After finding the best K candidate blocks of $\hat{x}_i$, for each candidate block $\hat{x}_i^{(k)}$ ($0 \leq k \leq K$), we separately apply each of the four *saliency reduction operators* $g_j(.)$ ($1 \leq j \leq 4$) as described in Section 5.3, in an attempt to lower the saliency value of the candidate block *without* increasing its l_2-norm error with respect to the thumbnail block $\tilde{y}_i$. More specifically, we use the following procedure for each lost MB $\hat{x}_i$:

1. Set $k \leftarrow 1$.
2. Set $j \leftarrow 1$.
3. Perform the j-th operator $g_j(.)$ on the k-th candidate block: $\hat{x}_i'^{(k)} \leftarrow g_j(\hat{x}_i^{(k)})$.
4. Given the saliency-reduced $\hat{x}_i'^{(k)}$, project $\hat{x}_i'^{(k)}$ onto the thumbnail $\tilde{y}_i$ using the method described in Section 5.2.
5. If the objective value (2) of the new $\hat{x}_i'^{(k)}$ obtained after step 4 is smaller than the smallest already-known objective function value, replace $\hat{x}_i^{(k)}$ with $\hat{x}_i'^{(k)}$ and go to step 3. Otherwise, go to step 6.

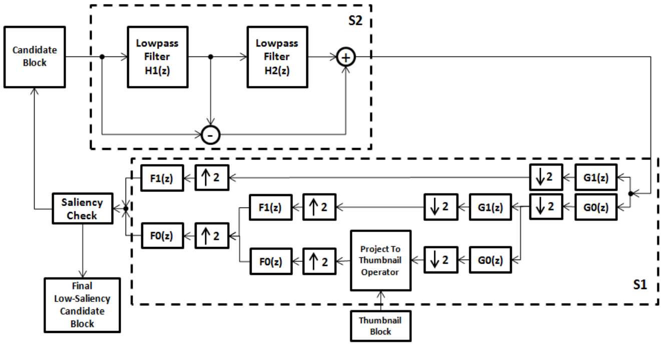


Fig. 2. The block-diagram of the proposed system. S1 shows the proposed system for projecting an input saliency-reduced candidate block to the thumbnail block. S2 is one of the four proposed saliency reduction operators, which is the notch filter in this figure.

6. Set $j \leftarrow j+1$. If $j \leq 4$, fetch the original candidate MB $\hat{x}_i^{(k)}$ again, and go to step 3. Otherwise, go to step 7.
7. Set $k \leftarrow k+1$. If $k \leq K$, go to step 2, otherwise to step 8.
8. Replace $\hat{x}_i$ with the best $\hat{x}_i^{(k)}$ whose objective function value is the smallest.

The above procedure is applied on all the lost MBs (i.e., $\hat{x}_i$'s) in a raster-scan order. Each time the visual saliency of $\hat{x}_i'^{(k)}$ needs to be calculated, we construct an adaptive window around the current MB whose top-left corner is set to the top-left corner of the video frame, and its bottom-right corner is set to the bottom-right corner of the current MB. The saliency of $\hat{x}_i'^{(k)}$ is then computed just within this adaptive window. Note that the adaptive window covers all correctly-decoded (or previously-concealed) MBs *plus* the current MB. A block-diagram of the proposed method is depicted in Fig. 2. The operation of each part of the proposed system is described next.

5.2. Projection onto the Thumbnail

In order to project a saliency-reduced candidate block $\hat{x}_i'^{(k)}$ to the thumbnail block $\tilde{y}_i$ as described in the algorithm in Section 5.1, we first down-sample $\hat{x}_i'^{(k)}$ by the same down-sampling factor used to generate the original thumbnail, by using a 4-tap conjugate wavelet filter bank [14], shown as subsystem S1 in Fig. 2. In this figure, $G_0(z)$ (low-pass) and $G_1(z)$ (high-pass) are the 4-tap analysis filters, and $F_0(z)$ and $F_1(z)$ are their corresponding conjugate synthesis filters [14]. After that, we compute the DCT coefficients of the coarsest low-frequency band. The projection to the thumbnail block is then accomplished by moving the 16×16 DCT coefficients of the coarsest band that are outside the designated quantization bin of the thumbnail block to the closest boundary of their respective quantization bins. The conjugate wavelet filter bank allows us to recover the exact $\hat{x}_i'^{(k)}$ when the low-frequency content of $\hat{x}_i'^{(k)}$ is already in good match with thumbnail $\tilde{y}_i$ due to its perfect reconstruction property. This operation is preformed on both the luma and chroma channels separately. This ensures that the low-frequency content of the new candidate block in all the channels remains in good match (in the l_2 norm sense) with the correctly-received thumbnail block.

5.3. Saliency-Reduction Operators

We next describe our proposed candidate operators $g_j(.)$, $1 \leq j \leq 4$, for reducing the saliency of a candidate MB $\hat{x}_i'^{(k)}$. These operators

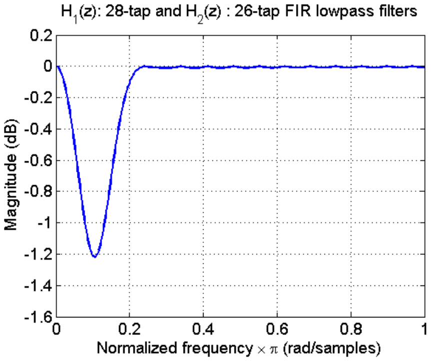


Fig. 3. The frequency response of the proposed notch filter.

are as follows: $g_1(.)$: notch filter, $g_2(.)$: frequency outlier filter, $g_3(.)$: intensity and color contrast reduction operator, and $g_4(.)$: deblocking filter.

5.3.1. *Notch Filter*

As mentioned in Section 4, the IKN model uses the frequency content of the input image in the range $[\pi/256 - \pi/16]$. Therefore, one simple way to reduce the saliency of a specific region (e.g., a candidate MB $\hat{x}_i^{(k)}$) is to reduce the strength of the signal in the aforementioned normalized frequency range. To achieve this goal, we propose to use a simple FIR notch filter [14] depicted as sub-system S2 in Fig. 2. This filter is composed of two low-pass filters whose transfer functions in the z-domain are denoted by $H_1(z)$ and $H_2(z)$ in Fig. 2, respectively. $H_1(z)$ is a lowpass FIR filter with 28 taps for the luminance channel and 14 taps for the chrominance channel. Similarly, $H_2(z)$ is a lowpass FIR filter with 26 taps for the luminance channel and 12 taps for the chrominance channel. The normalized cut-off frequency of the filters was set to $\pi/30$ (the center of the mentioned frequency range). These filters were designed by a standard window-based filter design method (Hamming method) [14] to achieve a normalized gain of -6 dB at the cut-off frequency, and a stop-band attenuation of about 50 dB. The frequency response of the obtained notch filter for the luminance channel is depicted in Fig. 3. As seen from this figure, the proposed notch filter has a very mild attenuation of about 1.2 dB at frequencies around $\pi/30$, which allows the system to reduce saliency slowly in each iteration of the proposed method. Note that any other filter design method can also be used here to obtain an appropriate notch filter, possibly with only one single FIR filter.

5.3.2. *Frequency Outlier Filter*

The second saliency-reduction operator we design is the *frequency outlier filter*. In [11], it was observed that since spatial frequencies in natural images follow an exponential decay in the power spectrum, a frequency component that does not follow the natural decay will be visually salient.

In our implementation, to lower the visual saliency of a candidate MB $\hat{x}_i^{(k)}$, we try to eliminate any potential frequency outlier in the candidate MB by comparing the frequency content of the candidate MB with the frequency content of its neighboring MBs. In order to achieve this goal, we first compute the DCT coefficients of the adjacent MBs in the 4-connected neighborhood around $\hat{x}_i^{(k)}$. After that, we compute an upper and lower bound for each DCT coefficient of $\hat{x}_i^{(k)}$ as follows:

$$u_{mn}^{upper} = \max(|u_{mn}^{top}|, |u_{mn}^{left}|, |u_{mn}^{right}|, |u_{mn}^{bottom}|), \tag{3}$$

$$u_{mn}^{lower} = \min(|u_{mn}^{top}|, |u_{mn}^{left}|, |u_{mn}^{right}|, |u_{mn}^{bottom}|), \tag{4}$$

where $u_{mn}^{top}, u_{mn}^{left}, u_{mn}^{right}$ and u_{mn}^{bottom} denote the DCT coefficients at frequency band (m, n), $0 \leq m, n \leq 15$, of the top, left, right, and bottom neighbors, respectively. Each DCT coefficient of the current MB is then clipped based on the computed upper and lower bounds as follows

$$u_{mn}^{*} = sign(u_{mn}) \times \begin{cases} u_{mn}^{upper} & \text{if } |u_{mn}| > u_{mn}^{upper}, \\ u_{mn}^{lower}, & \text{if } |u_{mn}| < u_{mn}^{lower}, \\ |u_{mn}|, & \text{otherwise}, \end{cases} \tag{5}$$

where u_{mn} is the original DCT coefficient of $\hat{x}_i^{(k)}$, and u_{mn}^{*} is the new DCT coefficient of $\hat{x}_i^{(k)}$ at frequency band (m, n). In our experiments, we applied this operator separately on the luminance channel (Y) as well as the two chroma channels (Cb and Cr) of $\hat{x}_i^{(k)}$.

5.3.3. *Intensity and Color Contrast Reduction*

Two of the important low-level features competing for visual attention in the IKN model are intensity and color contrast [8]. In [10], it was observed that if intensity and/or color contrast of a particular spatial region is enhanced, then the visual saliency as computed by the IKN model can be increased. For this purpose, [10] proposed to change the RGB components of each pixel within the desired region as follows:

$$\alpha_{pq}^{*} = \alpha_{pq} + w_{pq} V_{\alpha_{pq}}, \tag{6}$$

where α_{pq} denotes an RGB component ($\alpha = (R, G, B)$) of the pixel at location (p, q), α_{pq}^{*} denotes the updated α_{pq}, w_{pq} is a normalized positive weight factor, which is proportional to the saliency of the pixel at location (p, q), $V_{\alpha_{pq}}$ is a point variation factor, which reflects how much a feature influences the saliency of the pixel at location (p, q), and is computed by backtracking the saliency computation procedure in the IKN model. More details about this method can be found in [10].

Here, we would like to perform the opposite: reduce the intensity and/or color contrast of the candidate MB $\hat{x}_i^{(k)}$ so that its saliency value is decreased. For this purpose, we just negate the value of w_{pq} in (6), and apply (6) to all pixels within the candidate MB iteratively until its saliency cannot be decreased anymore.

5.3.4. *Deblocking Filter*

Another candidate operator for reducing saliency of a candidate block is the H.264/AVC deblocking filter proposed in [15]. It was shown in [15] that this simple deblocking filter can efficiently reduce the strength of blocking artifacts, which are usually attention grabbing especially at low bit rates. In our experiments, we found that this filter can reduce saliency of the concealed MBs as well. Hence, we propose this simple filter as a candidate saliency reduction operator.

Note that the strength of the H.264/AVC deblocking filter can be adjusted adaptively based on the quantization parameter of each MB as well as its coding mode [15]. Moreover, a separate boundary strength value (between 0 and 4) can be assigned to every edge between two 4×4 sub-blocks within a MB so as to be able to control the strength of the filter with a finer resolution. In H.264/AVC, these strength values are computed based on the coding mode of each MB. In our proposed method, however, we set all of the aforementioned boundary strength values to a small value (2 in our experiments), and we apply the deblocking filter several times on the candidate block until its saliency cannot be decreased anymore.

6. EXPERIMENTS

6.1. Experimental Setup

In our experiments, we used eight standard 30 frames per second (fps) sequences: *Bus* (CIF), *Crew* (CIF & VGA), *Football* (SIF), *Stefan* (SIF), *Soccer* (VGA), *Race Horses* (VGA), and *Keiba* (VGA) to test our proposed method. *Bus*, *Crew* (VGA), *Soccer*, *Race Horses*, and *Keiba* were 150 frames long, *Football* was 215 frames, and the other two sequences were each 300 frames long. CIF/SIF sequences were encoded at 700 kbps, while VGA sequences were encoded at 1400 kbps using the H.264/AVC JM 18.0 reference software [16], with the GOP structure IPPP. The thumbnail videos were created by downsampling their corresponding 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 to 10, K to 5, while the value of λ was experimentally set to 22. The IKN model with an extra motion and flicker channel [13] was utilized for the saliency computation, and the saliency values were between zero and one.

In order to find the most salient regions or ROIs, we first computed the saliency map of each video frame of each sequence. The saliency map of each frame was then binarized based on the 75-percentile of the saliency map of that frame. MBs with saliency above the 75-percentile threshold were considered as ROIs.

To simulate a real video streaming scenario with RECAP as its error control mechanism, a video frame was selected randomly, and its MBs in non-ROI parts were dropped randomly based on a two-state Gilbert model [17] at four different average loss rates (2%, 5%, 10%, 20%, and 30%) with an average burst loss length of 8. The corrupted frame was then concealed using both the original RECAP algorithm and our proposed method based on a correctly-received reference frame, which was assumed to be either 5 or 10 frames away. In practice, the distance between the concealed and reference frame is random. We used 5 and 10 simply as representative test values. This scheme was performed on about 30% of the total frames (randomly chosen) to get a loss-corrupted video out of each video sequence.

Fig. 4 shows an illustration of the visual quality of our proposed method compared to the original RECAP method for *Crew*. One can easily see that our method is able to improve the visual quality of the concealed frames compared to the original RECAP method.

6.2. PSNR Comparison

To show objective quality improvement of our scheme over the original RECAP (thus showing that low-saliency prior does provide correct side information to resolve the ambiguity in the replacement block search problem (2)), we constructed Table 1, showing the average PSNR (luma) improvement of our proposed method over the original RECAP algorithm at two reference frame distances $d = 5$ and $d = 10$. The average amount of saliency reduction (computed by the IKN model with an extra motion and flicker channel over all lost MBs, and averaged over both $d = 5$ and $d = 10$) brought by our proposed method is also mentioned in this table. These frame-level PSNR values have been computed at the aforementioned average loss rates, and only corrupted frames were considered for computing the average PSNR values. As seen from this table, the proposed method is able to improve the PSNR of the concealed frames compared to the RECAP method by up to 3.2dB, with an average saliency reduction amount of about 9%. The PSNR gains are larger at larger values of d.

Table 1. Average PSNR (dB) and saliency reduction amount achieved by the proposed method over RECAP at $d = 10$ and $d = 5$.

PSNR (RECAP : Proposed)

CIF & SIF Sequences	*Bus*	*Crew*	*Football*	*Stefan*
$d = 10$	28.8 : 30.0	26.6 : 29.8	24.1 : 25.9	26.3 : 27.2
$d = 5$	29.4 : 30.3	27.2 : 30.0	24.8 : 26.3	27.1 : 27.7
Average Saliency Reduction	10%	19%	12%	9%
VGA Sequences	*Soccer*	*Crew*	*Race Horses*	*Keiba*
$d = 10$	28.6 : 29.5	27.9 : 30.1	24.5 : 25.3	27.3 : 27.8
$d = 5$	29.2 : 29.9	28.5 : 30.3	25.1 : 25.5	27.9 : 28.1
Average Saliency Reduction	7%	13%	6%	4%

6.3. Subjective Testing

Since our proposed method aims at reducing the saliency of concealed regions, we performed a subjective test on CIF and SIF sequences to verify the improvement in subjective quality. In our experiment, a Two Alternative Forced Choice (2AFC) method [18] was used to compare subjective video quality. In 2AFC, the participant is asked to make a choice between two alternatives, in our case the original RECAP method and our proposed method. 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) [19].

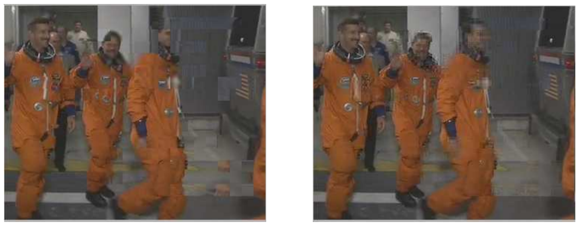
In each trial, participants were looking at two side-by-side videos (in the same vertical position, and separated by 1 cm horizontally) on a mid-gray background. Each video pair was shown for 10 seconds as recommended by ITU-R BT.500 [19]. After this presentation, a mid-gray blank screen was shown for 5 seconds. During this period, participants were asked to indicate on an answer sheet, which of the two videos looks better (Left or Right). They were asked to answer either Left or Right for each video pair, regardless of how certain they were of their response. Participants did not know which video was obtained by our method and which one was obtained by the RECAP method. Randomly chosen half of the trials had the video produced by our method on the left side of the screen and the other half on the right side, in order to counteract side bias in the responses. This gave a total of $4 \times 5 \times 2 = 40$ trials.

The experiment was run in a quiet room with 17 participants (all male except one, and of age between 18 and 30). All participants had normal or corrected to normal vision. A 24-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 actual height of the displayed videos on the screen was 87 millimeters. The illumination in the room was in the range 280-300 Lux. The distance between the monitor and the subjects was fixed at 70 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 are shown in Table 2, where we indicate the number of responses that showed preference for the original RECAP method and the proposed method at all of the tested average loss rates. We used the two-sided chi-square χ^2 test [20] to examine the statistical significance of the results. The null hypothesis is that there is no preference for either the RECAP method or the proposed method. Under this hypothesis, the expected number of votes is 17 for each method. The p-value [20] is also indicated in the table. In experimental sciences, as a rule of thumb, the null hypothesis is rejected when $p < 0.05$. When this happens in Table 2, it means that the

Table 2. Comparing the proposed method with the RECAP method based on the subjective results at 5 different average loss rates.

Loss Rate	Method	*Bus*	*Crew*	*Football*	*Stefan*
2%	RECAP	7	4	9	10
	Proposed Method	27	30	25	24
	p-value	0.0006	0.0001	0.0061	0.0164
5%	RECAP	4	3	7	9
	Proposed Method	30	31	27	25
	p-value	0.0001	0.0001	0.0006	0.0061
10%	RECAP	7	3	10	8
	Proposed Method	27	31	24	26
	p-value	0.0006	0.0001	0.0164	0.0020
20%	RECAP	8	8	11	7
	Proposed Method	26	26	23	27
	p-value	0.0020	0.0020	0.0396	0.0006
30%	RECAP	8	10	11	10
	Proposed Method	26	24	23	24
	p-value	0.0020	0.0164	0.0396	0.0164



(a) (b)

Fig. 4. Comparing the visual quality of the original RECAP method (left) with the proposed method (right) on *Crew*.

two methods 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.

As seen in Table 2, in all of the 40 trials the p-value is smaller than 0.05, which indicates that subjects showed a statistically significant preference for our proposed method. 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 (526 vs. 154 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.

Regarding the computational complexity, we emphasize that the main goal of our paper is to investigate the potential gain of using a low-saliency prior for error concealment in ROI-based UEP video streaming systems. Therefore, we did not optimize the proposed method for speed, which would generally be application- and platform-dependent in practice. We note that two of the four proposed saliency-reduction operators (i.e., the deblocking filter and the notch filter) have previously been implemented efficiently by others in other contexts, and none requires exponential running time. The other two operators can also be implemented efficiently. Further, in practice, loss happens only occasionally, hence the computation required for our method is needed only occasionally. Finally, we found experimentally that usually a few iterations (5-15) of the proposed method is sufficient to acquire acceptable results.

7. 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. In doing so, first, low saliency provides the right side information in ROI-based UEP video streaming systems for client to identify correct replacement blocks for concealment, and second, it reduces viewer's visual attention on the loss-stricken spatial 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 3.2dB in PSNR) and subjectively.

8. REFERENCES

[1] C. Yeo, W. t. Tan, and D. Mukherjee, "Receiver error concealment using acknowledge preview (RECAP)–an approach to resilient video streaming," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, April 2009.

[2] "Cisco visual networking index: Forecast and methodology 2010-2015," `http://www.cisco.com/`.

[3] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multi-view imaging and 3DTV," in *IEEE Signal Processing Magazine*, November 2007, vol. 24, no.6.

[4] G.-M. Muntean, G. Ghinea, and T.N. Sheehan, "Region of interest-based adaptive multimedia streaming scheme," in *IEEE Transactions on Broadcasting*, June 2008, vol. 54, no.2, pp. 296–303.

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

[6] N. Bruce and P. Kornprobst, "Region-of-interest intra prediction for H.264/AVC error resilience," in *IEEE International Conference on Image Processing*, Cairo, Egypt, November 2009.

[7] Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model," in *IEEE Transactions on Circuits and Systems for Video Technology*, June 2010, vol. 20, no.6, pp. 806–819.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, November 1998, vol. 20, no.11, pp. 1254–1259.

[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," in *IEEE Transactions on Multimedia*, January 2008, vol. 10, no.1, pp. 2–15.

[10] A. Hagiwara, A. Sugimoto, and K. Kawamoto, "Saliency-based image editing for guiding visual attention," in *1st International Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction*, Beijing, China, September 2011.

[11] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, June 2007.

[12] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami Beach, FL, June 2009.

[13] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," in *IEEE Transactions on Image Processing*, October 2004, vol. 13, no.10, pp. 1304–1318.

[14] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, 1993.

[15] P. List, A. Joch, J. Lainema, G. Bjntegaard, and M. Karczewicz, "Adaptive deblocking filter," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 614–619, July 2003.

[16] "The H.264/AVC JM reference software," [Online] Available: http://iphome.hhi.de/suehring/tml/.

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

[18] M. Taylor and C. Creelman, "PEST: Efficient estimates on probability functions," *J. Acoustical Society of America*, vol. 41, pp. 782–787, 1967.

[19] ITU-R, "Recommendation BT.500-8: Methodology for the subjective assessment of the quality of television pictures," Tech. Rep., ITU, 1998.

[20] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, 2007.