GAZE-DRIVEN VIDEO STREAMING WITH SALIENCY-BASED DUAL-STREAM SWITCHING

Yunlong Feng[o], Gene Cheung[#], Wai-tian Tan[*], Yusheng Ji[#]

[o] The Graduate University for Advanced Studies, [#] National Institute of Informatics,
[*] Hewlett-Packard Laboratories

ABSTRACT

The ability of a person to perceive image details falls precipitously with larger angle away from his visual focus. At any given bitrate, perceived visual quality can be improved by employing region-of-interest (ROI) coding, where higher encoding quality is judiciously applied only to regions close to a viewer's focal point. Straight-forward matching of viewer's focal point with ROI coding using a live encoder, however, is computation-intensive. In this paper, we propose a system that supports ROI coding without the need of a live encoder. The system is based on dynamic switching between two pre-encoded streams of the same content: one at high quality (HQ), and the other at mixed quality (MQ), where quality of a spatial region depends on its pre-computed visual saliency values. Distributed source coding (DSC) frames are periodically inserted to facilitate switching. Using a Hidden Markov Model (HMM) to model a viewer's temporal gaze movement, MQ stream is pre-encoded based on ROI coding to minimize the expected streaming rate, while keeping the probability of a viewer observing low quality (LQ) spatial regions below an application-specific ϵ. At stream time, the viewer's gaze locations are collected and transmitted to server for intelligent stream switching. In particular, server employs MQ stream only if: i) viewer's tracked gaze location falls inside the high-saliency regions, and ii) the probability that a viewer's gaze point will soon move outside high-saliency regions, computed using tracked gaze data and updated saliency values, is below ϵ. Experiments showed that video streaming rate can be reduced by up to 44%, and subjective quality is noticeably better than a competing scheme at the same rate where the entire video is encoded using equal quantization.

Index Terms— Region-of-Interest encoding, video streaming, visual saliency

1. INTRODUCTION

High-definition video is already a part of people's everyday life, from television to casual video captured in mobile phones. Similarly, streaming content for both live sources such as Skype and stored sources such as Netflix are increasingly moving towards high-definition and consumed on larger displays to keep up with user's expectation. This need to stream high-definition video to be consumed on large displays is both a curse and a blessing from an engineering perspective.

While high-definition streaming generally entails higher transmission cost, the use of large displays on both computers and televisions means that a viewer cannot simultaneously pay attention to all details being displayed. This is consistent with known observation that human's ability to perceive image details falls precipitously as function of the viewing angle away from his focal point of visual attention [1]. Therefore, one effective method to reduce transmission rate when viewing content on large displays is to employ a Region-of-Interest (ROI)-based approach [2, 3]. Specifically, if one can track a viewer's attention focal point in real-time, the server can encode a suitably sized spatial region (ROI) of the streaming video containing the viewer's focal point in high quality (HQ), and encode the other regions in low quality (LQ)[1]. However, there are two factors that can limit the practicality of such systems. First is the need of a computationally expensive live encoder for each viewer. Second is the need to react promptly to changes in visual focus with high probability.

In this paper, we propose a new ROI-based video streaming system that does not require real-time encoding, yet still reaps the benefit of bitrate reduction with low probability of visual quality degradation. We first pre-encode two bitstreams representing the same video content *a priori*. The first stream, called *HQ stream*, is encoded in HQ for entire video frames. The second stream, called *mixed quality (MQ) stream*, is encoded in two different qualities: spatial regions with *salient objects* (contiguous areas with per-pixel saliency values, pre-computed using a visual saliency model like [4], above a chosen threshold τ) are encoded in HQ, while other regions are encoded in LQ. In addition, to be discussed further in Section 3.2, Distributed Source Coding (DSC) frames [5] are periodically inserted into both streams to facilitate switching between the streams. Assume first that the server will switch from MQ to HQ stream at the next DSC frame boundary if viewer's gaze location falls outside the high-saliency regions (and vice versa). Using a Hidden Markov Model (HMM) [3] to model a viewer's temporal gaze movement, MQ is pre-

[1] Spatial regions outside ROI cannot be encoded using overly coarse quantization parameter (QP), so that the resulting coding artifacts draw unnecessary visual attention. We discuss selection of QPs in Section 3.5.

encoded to minimize the expected streaming rate, while keeping the probability of a viewer observing a LQ region to below an application-specific ϵ.

To achieve timely response to gaze movement, we employ a runtime gaze tracker at client with an eye-gaze prediction algorithm at the server. While it is not possible to guarantee that gaze prediction is always correct, prediction failures are generally less noticeable. This is because prediction failures usually take place during gaze redirection (called *saccade* in the literature), where it is known that human's ability to perceive details is decreased [6]. The probability of observing LQ regions due to switching delay associated with periodic insertion of DSC frames is hence contained in real-time as follows. The server updates the observation probabilities of saliency objects at DSC frame boundaries[2] given the real-time tracked gaze points, in order to recursively update probabilities of gaze evolution into future frames. The server chooses the MQ stream only if: i) viewer's latest tracked gaze location falls inside the high-saliency regions, and ii) the probability that a viewer's gaze moves and observes low-saliency regions before the next DSC boundary, calculated using updated object observation probabilities, is below ϵ.

Leveraging available software components for gaze tracking [7] and video encoding and streaming [8], we have built an operational video streaming system prototype that collects gaze data at 30Hz using a web camera and correspondingly makes intelligent stream-switching decisions during uninterrupted video streaming. Experiments show that video streaming rate can be decreased by up to 44% using our proposed system, and through an extensive subjective test that involved 20 human participants, we conclude with statistical confidence that the quality of our system is noticeably better than a competing scheme at the same streaming rate where the entire video is encoded using equal quantization.

The outline of the paper is as follows. We first overview related work in Section 2. We then discuss our system model in Section 3. We discuss how system parameters can be optimized and how optimal stream-switching decision can be made in Section 4 and 5. Finally, experimental results and conclusions are provided in Section 6 and 7, respectively.

2. RELATED WORK

Research in visual attention modeling has been very active in the last decades [4, 9], with numerous computational models proposed to identify spatial locations that attract the eye gaze. Most models compute a saliency map that values each pixel according to its visual saliency. Our goal here is not to propose new models of visual saliency maps, but to use saliency maps of video frames as per-pixel prior probability distribution of gaze location to estimate viewer's temporal eye movements in future frames. In this paper, we compute saliency maps using methodology in [4] based on a plausible model of bottom-up visual attention. This model offers good performance with reasonable computational cost. An existing implementation of the model is also available online.

In our previous work [3], we have designed a HMM to predict viewer's future gaze locations based on recent collected gaze data for real-time ROI-based video coding & streaming over networks with non-negligible delay. While using a similar HMM, our current work addresses a different problem: how real-time video encoding can be avoided altogether while still reaping the benefit of reduced rate in gaze-driven video streaming.

3. SYSTEM DESCRIPTION

We first overview our proposed gaze-driven video streaming system. We then describe the dual-stream frame structure used for stream-switching, and discuss how saliency objects are identified using visual saliency maps.

3.1. Streaming System Overview

Our proposed store-and-playback video streaming system employs two pre-encoded video streams with the same content in different qualities: HQ stream has all spatial regions encoded in HQ, while MQ stream only has visually salient regions encoded in HQ. DSC frames are inserted periodically every T frames to facilitate stream-switching depending on real-time tracked gaze locations. The system essentially switches to HQ stream when a viewer's gaze travels outside visually salient regions, and switches back to MQ stream when the server is confident that the viewer's gaze will remain in visually salient regions in the foreseeable future.

In theory, it is possible to set T small enough so that *zero* visual degradation is observed. This is because when human gaze shifts from one object of interest to another—a movement called *saccade*—the observer cannot perceive visual details until his vision has settled on the new object [6]. Hence if T is small enough that the server can switch from MQ stream to HQ stream before saccade has completed, the observer will always perceive HQ. Doing so would require a very small T, however, which is not practical given the non-negligible overhead of encoding stream-switching DSC frames. Hence, we take the alternative approach of limiting the probability of observing LQ regions to be below an application-specific ϵ instead.

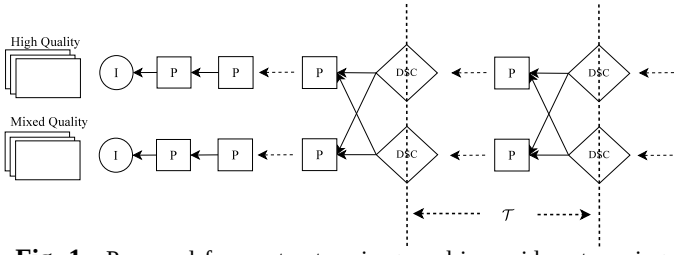


Fig. 1. Proposed frame structure in gaze-driven video streaming system. I-, P-, and DSC frames are denoted as circles, squares, and diamonds, respectively. DSC frames are inserted every T frames.

[2] A saliency map for a video frame can be interpreted as probability distribution function (PDF) of a viewer's per-pixel visual attention.

3.2. Dual-quality Frame Structure

We now describe the frame structure used to facilitate periodic switching between two pre-encoded bitstreams, shown in Fig. 1. HQ stream is encoded in HQ for entire video frames. MQ stream is encoded in two quality: spatial regions with per-pixel visual saliency values above a saliency threshold τ (ROI) are encoded in HQ, while the other regions are encoded in LQ. Frames in each stream are encoded in IPPP structure, with DSC frames [5] periodically inserted with period T frames to enable stream-switching at DSC frame boundary. More specifically, each DSC frame W^q_{nT} of instant nT, $n \in \mathcal{Z}^+$ and $q \in \{\mathrm{HQ}, \mathrm{MQ}\}$, is encoded with two predictor P-frames of previous instant $nT-1$ from the two streams, P^{HQ}_{nT-1} and P^{MQ}_{nT-1}. The reconstruction property of DSC frame guarantees that W^q_{nT} can be correctly decoded if *any one* of the predictor frames is available at decoder buffer as side information. Thus, a client can switch from P^{HQ}_{nT-1} in HQ stream or P^{MQ}_{nT-1} in MQ stream to W^q_{nT} in q stream at DSC boundary nT.

3.3. Identification of Saliency Object

We now describe how ROI is determined in MQ stream for pre-encoding in HQ. First, per-pixel visual saliency map for each video frame is computed using methodology in [4]. Saliency values of a spatial region roughly correspond to the amount of visual attention the region is likely to attract from viewers. Having computed visual saliency maps, we first normalize each one, so that the sum of saliency values in each map equals to 1. Then, spatially connected pixels with saliency value larger than a saliency threshold τ are grouped together as a *saliency object*. See Fig. 2 for examples of saliency objects in video frames. Saliency objects are encoded in HQ, as described earlier.

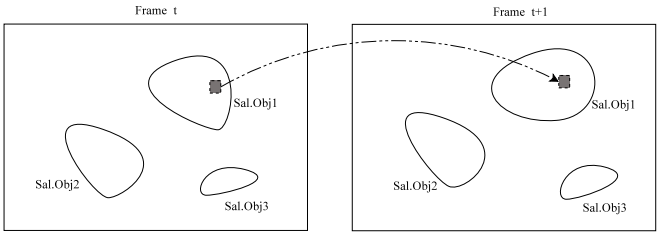


Fig. 2. Saliency objects in video frames, and how correspondence among them are found using motion estimation.

3.4. Temporal Correspondence of Saliency Objects

Having identified saliency objects in each frame, we can establish correspondence among saliency objects in consecutive frames using motion estimation (ME), commonly used in video coding standards like H.263 [10]. In details, for each $n \times n$ *target block* in a saliency object $o_{t+1,i}$ in frame $t+1$, we find the most similar *predictor block* in frame t, i.e., the block in frame t with corresponding RGB pixel values most similar to target block in frame $t+1$. If a sufficiently large fraction of target blocks of $o_{t+1,i}$ in frame $t+1$ map to blocks of the same object $o_{t,i'}$ in frame t, then we declare they are the same object. If no such object exists in previous map t, then we declare object $o_{t+1,i}$ to be a new object appearing for the first time in map $t+1$. As an example, in Fig. 2, we see that a block in object 1 in frame $t+1$ has found a matching block in object 1 in frame t.

3.5. QP Selection for MQ Encoding

Obviously, the coarser the QP used for encoding of spatial regions outside saliency objects (non-salient regions), the smaller the streaming rate. However, if non-salient regions are encoded using an overly coarse QP, then the resulting coding artifacts could potentially draw unnecessary visual attention, which is not desirable. To avoid this, we select the QPs for salient and non-salient regions in MQ stream as follows. QP for saliency objects in MQ stream and all spatial regions in HQ stream is first selected based on visual quality required for the given streaming application. Then, QP for the non-salient spatial regions in MQ stream is selected to be the coarsest possible, such that the difference between the visual saliency maps of frames in HQ stream and of frames in MQ stream, as measured by Kullback-Leibler (KL) divergence, remains no larger than a pre-defined threshold ψ. This ensures that the visual salient objects in the original frames are still the most salient regions, even after unequal quantization.

4. DUAL-STREAM CODING OPTIMIZATION

We are now ready to formulate our optimizations: i) how MQ stream is optimized during coding, and ii) how optimal stream-switching decision can be made at stream time with the benefit of real-time collected gaze data.

4.1. Hidden Markov Model for Gaze Movement

We first briefly discuss a simplified version of a previously proposed HMM [3] for viewer's temporal gaze movement during a video streaming session. The simplified HMM has two latent states: fixation, and saccade. State F (*fixation*) models the case when eye gaze is fixated at an object. State S (*saccade*) models the case where gaze rapidly moves from one fixation point to another. An HMM is Markovian in that the determination of state variable X_{n+1} at time $n+1$ (F or S) depends solely on the value of X_n of previous time n. In particular, given $X_n = i$, the probability of $X_{n+1} = j$ is represented by *state transition probability* α_{ij} of switching from state i to j. We assume that the speed at which a viewer's gaze moves from one object of interest to another is slower than the sampling rate of the gaze data (hence he must enter saccade state first before entering in fixation state again for observation of the new object). Our previous empirical results [3] showed this is approximately true.

Assuming stationar gaze movement statistics in a short video sequence, $\alpha_{i,j}$'s can be estimated using either collected eye gaze traces of test subjects [3], or off-line analysis of the video frames' saliency maps [11]. See [3] and [11] for details.

4.2. Saccade & Non-salient Observation Probabilities

Using the HMM, we can derive the *non-salient observation probability* e_t: the probability that a viewer will observe spatial regions different from the designated saliency objects in frame t during normal video playback. e_t's are crucial in deriving the *failure probability* ρ: the probability that a viewer will observe LQ spatial regions using our dual-stream system.

Let $p_{t,i}$ be the *saliency object probability* of a viewer observing a saliency object $o_{t,i}$, in frame t. $p_{t,i}$ can be computed simply as the sum of the per-pixel normalized saliency values within the saliency area. Let s_t be the *saccade probability* that a viewer is switching from one object of interest to another (in saccade state) in frame t. s_t and e_t are unknown for each frame t. To find these unknowns, we first derive two sets of equations in consecutive frames.

4.2.1. Total Probability Equations

First, we know that for each frame t, the sum of saliency object probabilities $p_{t,i}$'s, saccade probability s_t, and non-salient observation probability e_t equals to one due to the total probability theorem:

$$\sum_i p_{t,i} + s_t + e_t = 1 \tag{1}$$

(1) obviously holds true for probabilities in any frame.

4.2.2. Consistency Equations

We can write consistency equations for consecutive frames given the constructed HMM with estimated $\alpha_{i,j}$'s. Assume first that there are no new saliency objects in frame $t+1$; i.e., each object $o_{t+1,i}$ in frame $t+1$ has a corresponding object $o_{t,i'}$ in frame t. By the HMM, probability $p_{t+1,i}$ is the the sum of: i) saliency object probability $p_{t,i'}$ that a viewer watched object $o_{t,i'}$ in frame t multiplied by probability α_{FF} that the viewer stays in the same object $o_{t+1,i}$ in frame $t+1$, and ii) saccade probability s_t that a viewer was in transition in frame t and switches to object $o_{t+1,i}$ in frame $t+1$:

$$p_{t+1,i} = \alpha_{FF}\, p_{t,i'} + \alpha_{SF}\, s_t \left(\frac{p_{t+1,i}}{\sum_j p_{t+1,j} + e_{t+1}} \right) \tag{2}$$

Note that in (2), there is a scaling factor in the second term, indicating that only a proportional fraction of the saccade transition probability $\alpha_{SF}\ s_t$ will enter object $o_{t+1,i}$. (2) holds true for all saliency object probabilities $p_{t+1,i}$'s in frame $t+1$, as well as non-salient observation probability e_{t+1}. For saccade probability s_{t+1}, we can write it as a simple sum of: i) saccade probability s_t in frame t multiplied by the probability that the viewer stays in saccade state, and ii) the probability that a viewer first enters into saccade state at frame $t+1$:

$$s_{t+1} = \alpha_{SS}\, s_t + \alpha_{FS} \left(\sum_j p_{t,j} \ + \ e_t \right) \tag{3}$$

If frame $t+1$ has k saliency objects, then we have $k+2$ consistency equations for consecutive frame t and $t+1$. Together with the two total probability equations for two frames, we have a total of $k+4$ equations. However, we only have 4 unknowns: s_t, e_t, s_{t+1} and e_t. Hence in general, we have more equations than unknowns[3], and equations are not linear in unknown variables. To resolve this in a computationally efficient manner, we first identify the object $o_{t+1,i}$ with the largest saliency object probability $p_{t+1,i}$, and solve for s_t in (2) assuming e_{t+1} is 0. Given (1), e_t can then be computed. Finally, having computed s_t and e_t, s_{t+1} can then be computed using (3).

4.3. Dual-Stream Coding Optimization

We now describe how MQ stream coding is optimized. Assuming server switches from MQ to HQ stream when a viewer's gaze travels outside high-saliency regions (and vice versa), the objective is to minimize expected streaming rate while keeping the failure probability ρ below an application-specific value ϵ. There are two degrees of freedom in the optimization: i) saliency threshold τ used to define saliency objects in each frame t (as described in Section 3.3), and ii) DSC frame insertion period T.

4.3.1. Objective Function for MQ Stream Optimization

Given computed saccade and non-salient observation probabilities for each frame t as discussed in Section 4.2, the expected streaming rate R can be written simply. Let $|F_t^q|$ be the size of frame F_t^q of quality q and instant t. R is the sum of sizes of $N(T)$ segments, where each segment n of T frames (starts with a DSC frame) is HQ if viewer's gaze is outside high-salient regions with probability $s_{nT} + e_{nT}$, and MQ otherwise:

$$R = \sum_{n=1}^{N(T)} \left[(s_{nT} + e_{nT}) \sum_{k=0}^{T-1} |F_{nT+k}^{\text{HQ}}| + (1 - s_{nT} - e_{nT}) \sum_{k=0}^{T-1} |F_{nT+k}^{\text{MQ}}(\tau)| \right] \tag{4}$$

where each frame $F_t^{\text{MQ}}(\tau)$ in MQ stream depends on saliency threshold τ. Large τ means designated saliency objects are smaller, hence fewer pixels require HQ encoding and the resulting frame $F_t^{\text{MQ}}(\tau)$ is smaller.

4.3.2. Failure Probability Constraint

We now compute failure probability ρ. Failure happens if a viewer is observing a saliency object $o_{nT,i}$ at the DSC frame nT with probability $p_{nT,i}$, but drifts to a low-salient region *after* going through at least one saccade state in an intermediate frame. See Fig. 3 for an illustration. Note that ρ is *not* simply the average of non-salient observation probabilities e_t's, since here we are computing ρ assuming the viewer watches a saliency object in frame nT, while e_t's are computed unconditionally.

Nevertheless, we can derive the probabilities that a viewer is observing a saliency object, is in saccade, or is observing

[3]This is due to our assumption of stationarity of gaze statistics (roughly true for short video sequence), resulting in fixed $\alpha_{i,j}$'s for all frames.

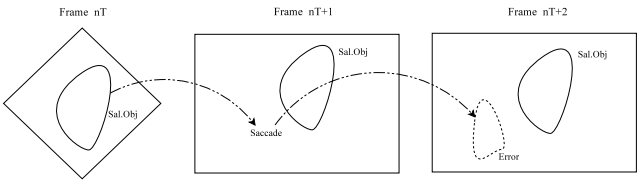


Fig. 3. Example where viewer observes LQ spatial region at frame $nT + 2$.

a low-saliency region, (P_{nT+k}, S_{nT+k} and E_{nT+k} respectively) similar to our consistency equations in Section 4.2.2. At the DSC frame nT, we have initial condition: $P_{nT} = 1$, $S_{nT} = E_{nT} = 0$. Probabilities of subsequent frames $nT + k$, $k = \{1, \ldots, T\}$, can be derived recursively:

$$\begin{aligned} P_{nT+k+1} &= \alpha_{FF} P_{nT+k} + \alpha_{SF} S_{nT+k} \left(\frac{\sum_i p_{nT+k+1,i}}{\sum_i p_{nT_k+1,i} + e_{nT+k+1}} \right) \\ E_{nT+k+1} &= \alpha_{FF} E_{nT+k} + \alpha_{SF} S_{nT+k} \left(\frac{e_{nT+k+1}}{\sum_i p_{nT_k+1,i} + e_{nT+k+1}} \right) \\ S_{nT+k+1} &= \alpha_{SS}\, S_{nT+k} + \alpha_{FS}\, (P_{nT+k} + E_{nT+k}) \end{aligned} \quad (5)$$

ρ is then just the average of E_{nT+k}'s for all k and n:

$$\rho = \frac{1}{N(T)} \sum_{n=1}^{N(T)} \frac{1}{T} \sum_{k=1}^{T} E_{nT+k} \quad (6)$$

4.3.3. Optimization Procedure

To minimize R while keeping $\rho < \epsilon$, we perform a greedy search: starting with the smallest saliency threshold τ and DSC insertion period T possible, we iteratively attempt to increase each one until the constraint $\rho < \epsilon$ can no longer be maintained.

5. REAL-TIME STREAM-SWITCHING OPTIMIZATION

During actual streaming, the viewer's eye gaze is tracked real-time by a normal web camera. The gaze data allow us to make intelligent stream-switching decisions as follows. At the DSC frame boundary $t = nT$, we know the viewer is observing a particular saliency object $o_{nT,i}$ with certainty (if viewer's gaze point is outside saliency objects, we transmit HQ stream). Thus, using initial condition $p_{nT,i} = 1$, $p_{nT,j} = s_{nT} = e_{nT} = 0, \forall j \neq i$, we can compute saliency object, saccade and non-salient observation probabilities for future frames using recursive equations similar to (5). Failure probability ρ until the next DSC frame can be computed similar to (6). Server will hence switch to MQ stream only if computed $\rho < \epsilon$.

6. EXPERIMENTATION

6.1. Experimental Setup

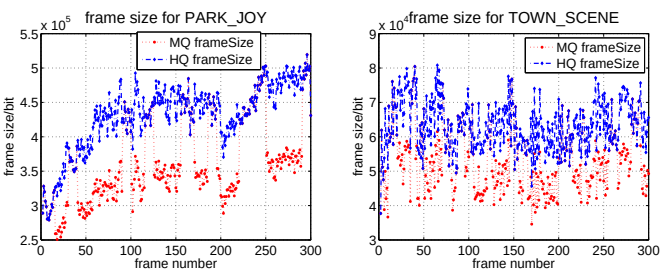
To demonstrate the effectiveness of our proposed system, we used two 300-frame HD video test sequences, `park_joy` and `town_scene`, at resolution 1280×720. Both videos have 30 frames per second (fps) playback speed. For simplicity, DSC-frame insertion period was fixed at $T = 5$, which from empirical evidence achieved a good tradeoff between streaming rate and the probability of observing LQ. HQ stream was encoded with quantization parameter (QP) set at 10, while MQ stream has low-saliency region encoded at a coarser (larger) QP value (QP_2) to be discussed later. A LQ stream was encoded at fixed QP with roughly the same bitrate as the competing scheme.

The experiments were conducted with a standard web camera running free eye-tracking software [7] in a quiet room with 20 subjects (12 males and 8 females, and of age between 21 and 40). All participants had normal or corrected to normal vision. A 24-inch Dell monitor with the native resolution 1920×1200 was used, with brightness and contrast set to 30% and 50%. The sequences were displayed in their original resolution, whose actual height on the screen was 196 millimeters. Each subject sat at a station facing a monitor with a fixed distance 55cm, and was asked to hold his/her head motionless.

In each trial, a subject was shown two videos back-to-back (with 3 seconds break in-between) at center of the screen. Each video lasted for 10 seconds as recommended by ITU-R BT.500 [12]. After these presentations, subject was asked to indicate which of the two videos looks better (First or Second). Each subject was familiarized with the task before the start of experiment with a short instruction. Two random combinations of two from HQ, MQ and LQ, using 4 different QP_2 for low-saliency region of MQ stream and equal quantization parameter for encoding of LQ stream, gave a total of $3 \times 4 \times 2 = 24$ pairs. Then, a Two Alternative Forced Choice method [13] was used to compare subjective video quality.

6.2. Experimental Results

During video playback, the viewer's gaze points were tracked and sent to server. The server then performed optimal stream-switching decisions at DSC frame boundaries. Fig. 4 shows example traces of how the system switched between HQ and MQ streams (red) as function of time. Also shown are bitrates if HQ streams were used at all frames (blue). It is observed that the performance is correlated with the character-



(a) trace for `park_joy` (b) trace for `town_scene`

Fig. 4. Frame size versus frame number for two video sequences when $QP_2 = 12$.

QP_2	KL div.	*MQ* : *HQ*	*MQ* : *LQ*	*HQ* : *LQ*
12	1.09E-08	23:17	35:5	33:7
		(0.343)	(2.1E-06)	(3.94E-05)
15	1.90E-08	17:23	38:2	40:0
		(0.343)	(1.25E-08)	(2.54E-10)
20	2.36E-08	17:23	38:2	39:1
		(0.343)	(1.25E-08)	(1.87E-09)
25	3.08E-08	11:29	38:2	38:2
		(0.0044)	(1.25E-08)	(1.25E-08)

Table 1. Number of viewer who prefer the schemes HQ, MQ and LQ, with corresponding p-value shown in parenthesis.

istics of test sequences: `park_joy` contains higher motion than `town_scene`.

The subjective testing results are given in Table 1, where we indicate the number of responses showing preference for HQ, MQ and LQ at different QP_2 values. We used the two-sided chi-square test [14] to examine the statistical significance of the results. The null hypothesis is that there is no preference for either two of HQ, MQ and LQ. Under this hypothesis, the expected number of votes is 20 for each method. The p-value [14] 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 1, it means that the two methods cannot be considered to have the same subjective quality, since one of them has obtained a statistically significantly higher number of votes.

As seen in Table 1, the subjects showed a statistically significant preference for our proposed MQ method and HQ over LQ, as the p-value is much smaller than 0.05 for all choices of QP_2. Furthermore, for a wide range of QP_2 between 12 and 20, the difference between our proposed method and HQ are statistically insignificant, with p-value significantly above 0.05. While more subjects prefer HQ at QP_2 of 15 and 20, it should be noted that an equal number prefer MQ over HQ at QP_2 of 12. This further indicates that such difference are not statistically significant. When QP_2 is increased to 25, the subjects show clear preference of HQ over MQ. The corresponding KL divergence is 3.08E-08. It is a subject of further study of whether KL divergence $\psi = 3.0$E-08 will ensure similar visual quality across sequences.

QP outside ROI	`park_joy`	`town_scene`
QP_2=12	20.41%	24.08%
QP_2=15	31.58%	31.10%
QP_2=20	44.14%	44.78%
QP_2=25	55.95%	59.10%

Table 2. Average bitrate reduction of MQ over HQ /10 times.

Table 2 shows the average bitrate reduction achieved by different values of QP_2. Over the large range of QP_2 between 12 and 25, meaningful bitrate savings between 20% and 60% can be obtained. In particular, at QP_2 of 20, savings of above 44% is realized without loss of visual quality that is statistically significant.

7. CONCLUSION

By only encoding spatial regions containing viewer's focal points of visual attention in high quality (HQ), ROI-based video streaming can reduce transmission rate without degrading perceived video quality. Unlike previous ROI-based systems that require real-time encoding, we present a system that switches between two pre-encoded streams based on real-time tracked gaze data. Streams are pre-encoded to minimize streaming rate while satisfying an application-specific quality requirement. Stream-switching decision is optimized based on tracked gaze data and real-time updated observation probabilities. Experiments using our constructed real-time streaming system show that bitrate can be reduced by up to 44% with test subjects noticing very little visual degradation.

8. REFERENCES

[1] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," in *Proceedings of the IEEE*, October 1993, vol. 81, no.10, pp. 1385–1422.

[2] Y. Liu, Z. G. Li, and Y. C. Soh, "Region-of-interest based resource allocation for conversational video communication of H.264/AVC," in *IEEE Transactions on Circuits and Systems for Video Technology*, January 2008, vol. 18, no.1, pp. 134–139.

[3] Y. Feng, G. Cheung, W. t. Tan, and Y. Ji, "Hidden Markov model for eye gaze prediction in networked video streaming," in *IEEE International Conference on Multimedia and Expo*, Barcelona, Spain, July 2011.

[4] 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.

[5] N.-M. Cheung, A. Ortega, and G. Cheung, "Distributed source coding techniques for interactive multiview video streaming," in *27th Picture Coding Symposium*, Chicago, IL, May 2009.

[6] L. Loschky and G. Wolverton, "How late can you update gaze-contingent multiresolution displays without detection?," in *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, December 2007, vol. 3, no.7.

[7] "Opengazer: open-source gaze tracker for ordinary webcams," http://www.inference.phy.cam.ac.uk/opengazer/.

[8] "FFmpeg," http://ffmpeg.org/.

[9] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," in *Vision Research*, September 2007, vol. 47, no.19, pp. 2483–2498.

[10] ITU-T Recommendation H.263, *Video Coding for Low Bitrate Communication*, February 1998.

[11] Y. Feng, G. Cheung, P. Le Callet, and Y. Ji, "Video attention deviation estimation using inter-frame visual saliency map analysis," in *IS&T/SPIE Visual Information Processing and Communication Conference*, Burlingame, CA, January 2012.

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

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

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