To appear *in ACM Transactions on Multimedia Computing, Communications and Applications Journal*, Vol. 5 Issue 1, February 2009.

Predictive Real-time Perceptual Compression based on Eye-gaze-position Analysis

OLEG V. KOMOGORTSEV
Department of Computer Science
Texas State University-San Marcos
ok11@txstate.edu
and
JAVED I. KHAN
Department of Computer Science
Kent State University
javed@kent.edu

This paper designs a real-time perceptual compression system (RTPCS) based on eye-gaze-position analysis. Our results indicate that the eye-gaze-position containment metric provides more efficient and effective evaluation of an RTPCS than the eye fixation containment. The presented RTPCS is designed for a network communication scenario with a feedback loop delay. The proposed RTPCS uses human visual system properties to compensate for the delay and to provide high ratios of multimedia compression.

Categories and Subject Descriptors: **I.6.4 [Simulation and Modeling]:** Model Validation and Analysis; **J.7 [Computers in Other Systems**]: Process control, Real time.
General Terms: Algorithms, Performance, Design, Reliability, Experimentation, Human Factors, Verification.
Additional Key Words and Phrases: real-time multimedia compression, human visual system.

---

1 INTRODUCTION

In this paper we present a design of a *Real-Time Perceptual Compression System* (RTPCS) based on eye-gaze-position analysis. The average eye-gaze-position containment is proposed as a new evaluation metric for an RTPCS performance. Additionally the proposed RTPCS is evaluated with respect to various parameterizations of the eye fixation detection.

An eye fixation is a type of eye movement that brings high acuity vision to the brain. Perceptual compression increases image quality around an eye fixation while reducing the image quality in the vision periphery in accordance with an eye sensitivity function.

Perceptual compression allows decreasing the bit-rate of the multimedia while preserving the same perceptual quality [3, 4, 5, 6].

Eye fixation analysis is the most common evaluation metric used today for eye tracking systems, but research literature still struggles with the exact method used for detecting eye fixations. Different eye fixation detection methods can lead to different RTPCS evaluation results. As a solution to this problem, we propose an eye-gaze-position containment metric. Our results show that such metric is more conservative and robust than an eye fixation metric.

Additionally, the design of the RTPCS described in this paper addresses the challenges of a networking scenario. Feedback loop delay associated with multimedia transmission presents the uncertainty about the location of future eye-gazes. We have designed a model which predicts future eye-gaze-position trace through previous eye movement analysis thus compensating for the delay.

The paper is organized in the following way. Section 2 presents a brief overview of related work, human visual system description and perceptual compression challenges. Section 3 outlines the design of our PRTCS. Section 4 describes experimental setup. Section 5 reports results including additional compression levels achieved by the proposed PRTCS. Section 6 presents a discussion on the system's limitations. Section 7 has the conclusion.

2 BACKGROUND AND OBJECTIVES

2.1 Previous work

There have been quite a few studies that investigated various aspects of perceptual compression. An excellent eye-tracking methodology book was written by Duchowski [25]. Research in perceptual compression field has mainly focused on the study of contrast sensitivity or spatial degradation models around an eye fixation and its impact on the perceived loss of quality by viewers [8, 10, 11, 12]. Geisler and Perry [13] presented pyramid coding and used a pointing device to identify the point of focus by a subject. Daly et. al. [14] presented an H.263/MPEG adaptive video compression scheme using face detection and visual eccentricity models. Bandwidth reduction of up to 50% was reported. For example, Daly [15] utilized a live eye-tracker to determine the maximum frequency and spatial sensitivity for HDTV displays with a fixed observer distance. Lee and Pattichis [6] discussed how to optimally control the bit-rate for an MPEG-4/H.263 stream for foveated encoding. Stelmach and Tam [17] have proposed to perceptually pre-

encode a video based on the viewing patterns formed by a group of people. Babcock et. al. [18] investigated various eye movement patterns and foveation placements during different tasks, making the conclusion that those placements gravitate toward faces and semantic features of an image.

A few researchers have worked on saliency maps and saccade target estimation in videos and 3D environments based on pre-computed image analysis [21, 22].

In our previous work, we have created an eye-speed-based scheme which looked at the perceptual compression for a single viewer [3] and multiple viewers [23] in a situation where the proposed perceptual compression model did not have access to the presented visual content. Later we developed several perceptual compression models which improve perceptual compression based on real-time scene analysis and content evaluation [24].

The work presented in this paper uniquely stands out from the previous research in terms of a design of a practical RTPCS that addresses some of the networking challenges, i.e., feedback loop delay and uneven eye-gaze-position sample arrival at an RTPCS. In this paper, we propose an eye-gaze-containment evaluation metric for the design of an RTPCS. Our results show that such a metric is more conservative than an eye fixation based metric. Our work builds on a single viewer approach presented in [3]. In this paper, we add to the past work by bringing the comparison between an eye-gaze-position and an eye fixation RTPCS evaluation, including more subjects, a wider test range of input parameters and a discussion that takes on the limitations of the proposed system.

2.2 Human visual system

There are three types of eye movements which are present when we look at multimedia: fixation, saccade, and smooth pursuit.

(i) Fixations: - "eye movement which stabilizes the retina over a stationary object of interest" [25]. Eye fixations are accompanied by drift, small involuntary saccades and tremor. A human's eye perceives the highest quality picture during an eye fixation. Eye fixations represent the areas of perceptual attention focus. Eye fixation duration usually ranges from 100 ms. to 600 ms. with eye velocity not exceeding 100 deg/s during a fixation. Usually ninety percent of the total viewing time in humans is spent in eye fixations [29].

(ii) Saccades: - "rapid eye movements used in repositioning the fovea to a new location in the visual environment" [25]. Saccade duration ranges from 10 ms. to 100 ms.

which renders the visual system blind during a saccade [20]. Saccade duration ranges from 30 ms. to 120 ms. with eye velocities going above 300 deg/s.

(iii) Smooth pursuits: - eye movements which develop when the eyes are tracking a moving visual target. It consists of these two components: a slowly varying motion component plus a saccadic component. This saccadic component occurs occasionally as a correction mechanism for the eye-gaze-position when the current eye-gaze-position is not accurate with respect to the moving object [19]. The slowly varying motion component keeps the retina stable over the moving object, and high quality visual data is perceived during this period.

The ability to perform perceptual compression comes from the anatomical properties of the human eye. The diameter of the eye's highest acuity, the fovea, extends only to 2 degrees. The parafovea, the next highest acuity zone, extends to about 4 to 5 degrees, and acuity drops off sharply beyond [2] that point.

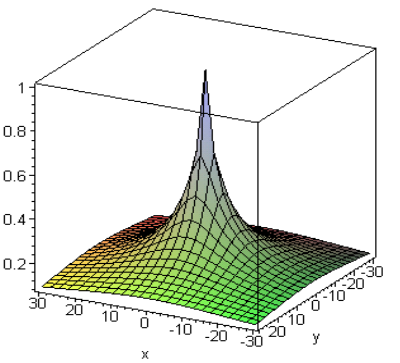


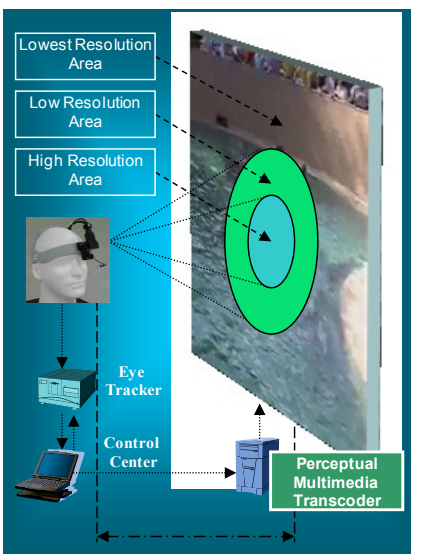
Figure 1. Visual sensitivity function.

Anatomical properties of the eye can be represented by a visual sensitivity function which allows us to perform perceptual compression of any multimedia in a form of image degradation from an eye fixation point to the pheriphery. In this paper we use a visual sensitivity function proposed by Daly and Ribas-Corbera [14]:

$$S(x,y) = \frac{1}{1 + ECC \cdot \theta_E(x,y)} \qquad (1)$$

Here, S is the eye visual sensitivity as a function of the image position (x,y), ECC is a constant (in this model ECC=0.24), and $\Theta_E(x,y)$ is the eccentricity in the visual angle. Figure 1 presents an example of S(x,y).

Within a specific RTPCS implementation, an eye sensitivity function has to be mapped to the spatial and temporal parameters of the selected codec.



2.3 Feedback loop delay

Previous research did not consider the impact of the feedback loop delay on perceptual compression. Feedback loop delay is the period of time between the instance the eye-gaze-position is detected by an eye-tracker and the moment when a perceptually compressed image is displayed. Figure 2 presents the feedback loop delay concept.

In the design of an RTPCS the delay compensation is important because future eye fixations should fall within the highest quality region of the image, preventing the viewer from noticing the compression artifacts.

Figure 2. Feedback loop delay during perceptual compression.

It is noteworthy that the properties of the multimedia transmission might change over time, thus increasing or decreasing the delay length. A typical network delay range is from 20 ms. to a few seconds. Saccades can move the eye position more than 10 degrees during that time while potentially placing a new eye fixation to the low visual quality region.

2.4 Objectives

Our first objective was to design an RTPCS that perceptually compresses a multimedia source using eye-gaze-position analysis in a network scenario with a delay. Our second objective was to make sure that the proposed eye-gaze-position metric is more conservative than an eye fixation-based metric.

3 REAL-TIME PERCEPTUAL COMPRESSION SYSTEM DESIGN

3.1 Perceptual Attention Focus Window

To compensate for the feedback loop delay we have created a concept of *Perceptual Attention Window* (W^{PAW}). The purpose of the W^{PAW} is to contain future eye fixations. Two parameters define the W^{PAW} - *Future Predicted Eye-Speed* (FPES) and feedback loop delay (T_d).

Our eyes move between the eye fixations using saccades. The acceleration, rotation, and deceleration involved in ballistic saccades are guided by the muscle dynamics and demonstrate stable behavior. The latency, direction of the gaze, and the eye fixation duration have been found to be highly dependent on the content of the media presented; and they are often hard to predict.

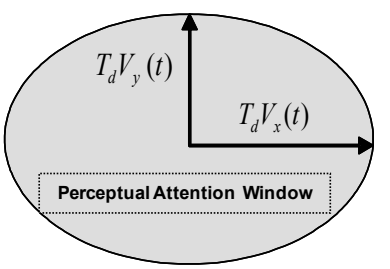


Figure 3. Perceptual Attention Window - W^{PAW}.

Therefore W^{PAW} is modeled as an ellipse, allowing the W^{PAW} boundaries to take any direction within the eye movement acceleration constraints. The size of the ellipse is proportional to the length of the feedback delay multiplied by the FPES. The FPES is broken into horizontal and vertical components which are represented by $V_x(t)$ and $V_y(t)$ values correspondingly. "t" is time when the FPES is calculated. Figure 3 presents a diagram of the W^{PAW}. Conceptually the W^{PAW} can be applied to any type of multimedia

The *Histogram Eye-Speed Analysis* (HESA) model is used for $V_x(t)$ and $V_y(t)$ calculation. The HESA description is presented in the Section 3.2.

The W^{PAW} transforms visual sensitivity function presented by (1) into a form:

$$S_t(x_{pix}, y_{pix}) = \begin{cases} 1, when \left(\sqrt{\left(\frac{V_y(t)\left(x_{pix} - x_{cen}^{PAW}(t)\right)}{V_x(t)} \right)^2 + \left(y_{pix} - y_{cen}^{PAW}(t)\right)^2} - V_y(t)T_d \right) < 0 \\ 1 \Big/ \left(1 + ECC\frac{180}{\pi} tan^{-1} \left(\frac{\sqrt{\left(\frac{V_y(t)\left(x_{pix} - x_{cen}^{PAW}(t)\right)}{V_x(t)} \right)^2 + \left(y_{pix} - y_{cen}^{PAW}(t)\right)^2} - V_y(t)T_d}{VD} \right) \right), \quad otherwise \end{cases} \tag{2}$$

where x_pix, y_pix are coordinates of every pixel of the image presented. $x\_cen^{PAW}(t)$ and $y\_cen^{PAW}(t)$ are the W^{PAW} center coordinates at the time instance "t". T_d is feedback delay length. VD is the distance between the viewer and the screen on which the multimedia content is presented. Note: all distances need to be converted to the pixel distances for this equation to be true.

Figure 4 presents a diagram of eye sensitivity function specified by (2). The peak point presented by the eye sensitivity function in Figure 1 becomes the ellipse of the W^{PAW}. That means that any point inside of the W^{PAW} has a sensitivity level equal to 1, and it will be encoded with the highest quality. The slope in Figure 4 is created by the eye visual sensitivity function represented by (1).

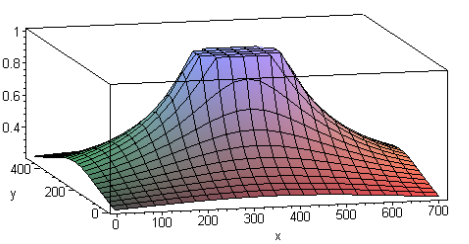


Figure 4. Eye Visual Sensitivity function combined with Perceptual Attention Window.

3.2 Histogram Eye-Speed Analysis

The intuitive goal of our algorithm was to assign an eye-speed value (*Running Frame Eye-Speed* - RFS) to the every video frame. Such an assignment should take care of cases when there are many eye-gaze-position samples detected for a frame and the cases when no eye-gaze-position samples are detected. RFS assignment should be conservative, i.e.,

in case of the high variance in terms of the eye-gaze-position coordinates and in case when the eye-tracker's sampling rate is much higher that the frame-rate of the RTPCS, the resulting RFS values would be higher, not lower. The resulting set of the RFS is evaluated through a histogram-based method to calculate Future Predicted Eye Speed.

Suppose there are n eye-gaze-positions sampled between frames F(t-1) and F(t) detected by the RTPCS. Each eye-gaze-position sample has (x_i,y_i) pixel coordinates on frame F(t). RFS for horizontal and vertical eye movement component is calculated as:

$$V_x^{RFS}(t) = \sum_{i=1}^{n-1} | x_{i+1}(t-T_F) - x_i(t-T_F) | \quad \textbf{(3)}$$

$$V_y^{RFS}(t) = \sum_{i=1}^{n-1} | y_{i+1}(t-T_F) - y_i(t-T_F) | \quad \textbf{(4)}$$

"n" is the number of eye-gaze-position samples detected for the frame "t". "n" can vary per frame due to the unpredictable system and network delays. T_F is the value of the feedback delay in the system measured in frames. $T_F=T_d/FrRt$. T_d is the value of the feedback delay in the system measured in seconds. FrRt is the RTPCS's current frame rate per second. Notation $x_i(t-T_F)$ and $y_i(t-T_F)$ shows that the eye-gaze-position samples that the RTPCS received for the frame F(t) are T_d seconds late. Thus delayed eye-gaze-positions are represented by coordinates $x_i(t-T_F)$ and $y_i(t-T_F)$, where $1 \le i \le n$, and n is the number of eye-gaze-position coordinates received by the RTPCS while the frame F(t) is being compressed. Such computation of the RFS will satisfy the requirements described at the beginning of this section. The HESA model uses $x_n(t-T_F)$ and $y_n(t-T_F)$ as coordinates of the W^{PAW} center for the frame F(t).

After each frame was assigned an RFS number, it would be necessary to select a boundary for the Future Predicted Eye Speed based on the history presented by the RFS set. The RFS set represents the "memory" of the previous eye movement behavior. The RFS "memory" is represented by a histogram which allows it to "cut off" the unnecessary high RFS values present mostly due to the saccades. As a result of the "cut off," future eye-gaze-positions representing a saccade will not be covered by the W^{PAW}, but such RTPCS performance is satisfactory due the fact that our eyes are blind during the saccades. Eye-gaze-position samples representing the eye fixations and the pursuits will be covered by the W^{PAW}.

The "cut off" parameter is represented by the *Target Eye-Gaze Containment* (TGC). The choice of this parameter depends on the amount of the saccades and noisy eye-gaze-position samples presented in the eye trace. The "memory" of the histogram created by the RFS values will be represented by the RFSs (*Running Frame Eye-Speed samples*)

parameter. Parameter RFSs represents the size of the set containing a specific number of the most recent RFS samples.

Mathematically the HESA model works as follows:

Two RFS sample sets are created: $\{V_x^R(t-RFSs),..,V_x^R(t)\}$ and $\{V_y^R(t-RFSs),..,V_y^R(t)\}$. The TGC parameter creates a percentile "cut off" boundary q in each RFS set: $q = \left\lfloor \frac{TGC}{100} \right\rfloor RFSs$. After this step, a Randomized-Select algorithm described in Cormen et. al. [1] is used to calculate the FPES values.

$$V_x(t) = RndSel(V_x^{RFS}, t-RFSs, t, q) \quad \textbf{(5)}$$

$$V_y(t) = RndSel(V_y^{RFS}, t-RFSs, t, q) \quad \textbf{(6)}$$

The Randomized-Select algorithm returns the value of q[th] smallest element of the input array and it runs in O(RFSs) time. The value of the q[th] smallest element represents

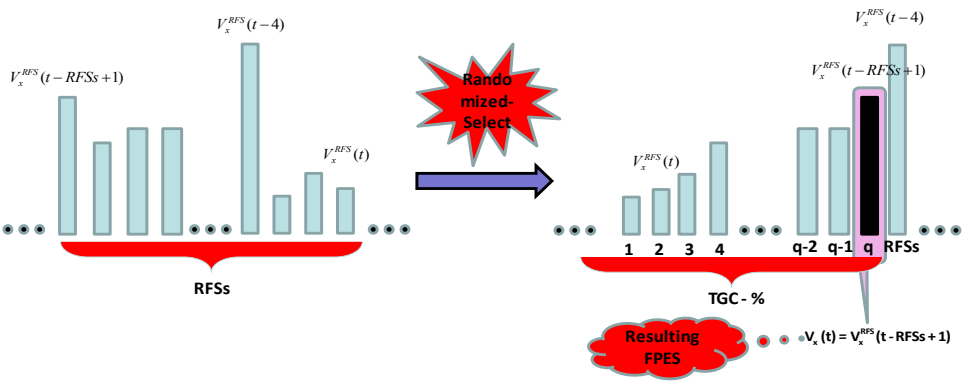


Figure 5. Example of a horizontal FPES calculation.

the Future Predicted Eye Speed necessary to satisfy the goals of the HESA model. The percentage of the RFS samples between the smallest element and the q[th] element is less or equal to the TGC value. An example of the FPES calculation by the HESA model is presented in Figure 5.

4 EXPERIMENT

4.1 Equipment

The proposed RTPCS was evaluated using an MPEG-2 transcoder [28] integrated with Applied Science Laboratories eye-tracker model 504. ASL 504 has the following characteristics: accuracy - spatial error between true eye position and computed measurement is less than 1 degree; precision - better than 0.5 degree; eye-gaze-position scanning rate – 60Hz. That model of the eye-tracker compensates for small head movements within a few inches so the subject's enforced head stabilization was not

required. Nevertheless, during the experiments, every subject was asked to hold his/her head still. Before running each experiment, the eye-tracking equipment was calibrated for the subject and checked for the calibration accuracy; and if one of the calibration points was "off", then the calibration procedure was repeated for that point.

4.2 Eye Fixation Detection Algorithm

Surprisingly, there is no firm definition for an eye fixation. It should be noted that from a practical point of view, an eye fixation is less a physiological quantity than a method for categorizing sections of a data stream. Sensible selection of criteria depends on the experimental goal and the characteristics of the measurement as well as the underlying physiology. There are quite a few different algorithms in the literature for detecting eye fixations [25], all of which represent logical strategies. Processing the same data with

Figure 6. Example of the perceptual (left side) and uniformly (right side) compressed frames for the "Car", "Shamu", "Airplanes" videos. Target bit-rate is 1Mb/s.

different algorithms or different parameters for a given algorithm easily results in a different number of eye fixations and different sets of eye fixation start and stop times and positions.

The algorithm that we used for eye fixation detection falls in the category that Duchowski [25] labels "dwell-time eye fixation detection". The full description of the eye fixation detection algorithm is presented in the ASL eye-tracker manual [26]. To detect an eye fixation, this algorithm looks for a specified time period β (minimum fixation duration) when eye-gaze-position samples within this time period have a standard deviation of no more than γ degrees of the visual angle. γ parameter was set to 0.5 degrees (0.5 is the maximum amplitude of the involuntary saccades within an eye fixation [7]). β parameter was varied between 100 ms. and 150 ms. A 100 ms. duration is recommended by eye-tracker manufacturers [16, 26], and 150 ms. is the time duration that is suggested by eye-tracking research literature [25]. In our experiments, both values of β are used to evaluate HESA based RTPCS.

4.3 Test Multimedia Content

Human eye movements are highly dependent on the visual content. Some types of scenes inherently offer more opportunity for compression and some offer less. Any multimedia compression algorithm should continuously analyze the complexity of a scene and provide the best performance possible. Unfortunately, there is no easy or agreed means of measuring the complexity of the content. To select our test bed clips, we have looked at several examples, each offering different combinations of subjective complexities. In this paper, we consider three representative cases. Each selected video clip presents different content challenges to our RTPCS. Below are rough subjective complexity descriptions for each video clip:

Car: This video shows a moving car. It was taken from a security camera viewpoint in a university parking lot. The visible size of the car was approximately one fifth of the screen. The car was moving slowly, allowing the subject to develop smooth pursuit eye-movement (our assumption). Sometimes there are smaller objects such as pedestrians and other cars which appear briefly in the background, but mainly this video's background is stationary.

Shamu: This video captures an evening performance of Shamu at a Sea World during the nighttime under a tracking spotlight. This video consists of several moving objects: Shamu, the trainer, and the crowd. Each object is moving at different speeds during various periods of time. The interesting aspect of this video is that a subject can

concentrate on different objects, and it would result in a variety of eye movements: fixations, saccades, and smooth pursuit. The background of the video was constantly moving due to the fact that the camera was trying to follow a moving Shamu. Such an environment suits the goal of challenging our algorithm to deal with different types of eye movements. The fact that the clip was taken during the night provides an interesting aspect of the video perception behavior by a subject. The snapshot is presented in Figure 6a.

Airplanes: This video depicts formation flying of supersonic planes, rapidly changing their flying speeds. It was from a performance of the Blue Angels over Lake Erie. The number of planes varies from one to five during the clip. The scene recording camera movements were rapid zoom and panning. This video provided a challenge to the human visual system – the capturing camera moves unexpectedly, making the HVS "overshoot" the airplane location. This behavior challenges the HESA model to build a compact W^{PAW} to contain the unexpected eye shifts. The background of this video was in constant motion and presented a blue sky.

Figure 6 shows an example of the perceptually and the uniformly compressed frames using the same bit-rate. It is possible to see that the areas where a viewer is looking are blurry in uniformly compressed frames but have a much better quality at perceptually compressed frames.

All three videos had a resolution of 720x480 pixels, presented with the frame-rate of 30fps, and were between 1 and 2 minutes long. The original and perceptually compressed video clips are available at our website [27].

4.4 Participants

Three subjects have participated in the evaluation experiments. Each of them had normal or corrected to normal vision. The subjects were not aware of the video content before the experiments and were asked to look at the presented content in any way they wanted. This type of setup is called free-viewing in eye-tracking literature. Test videos were presented on the screen of an 18 inch LCD monitor. The distance between the subjects' eyes and the monitor surface was 43 inches. The size of the screen measured 261x241 in eye-tracker units and had a pixel resolution of 1280x1024.

4.5 Raw eye-gaze-position data filtering

An eye position sample was classified as noisy when the eye tracker failed to measure the eye position coordinates for that sample. The failure to identify eye position

coordinates usually happens due to the subject's blinking, jerky head movements, changes in the content's lighting, excessive wetting of the eye, and squinting. The coordinates of each noisy eye position sample were replaced with the coordinates of the previous successfully measured eye position sample.

4.6 Evaluation parameters

The HESA based RTPCS is validated through the *Average Eye Fixation Containment, the Average Eye-Gaze-Position Containment, and the Average Perceptual Resolution Gain*.

4.7 Eye-gaze and eye fixation containment

The *Average Eye-Gaze-Position Containment* (AEGC) is the main design parameter for our system. The AEGC reports how many raw eye-gaze-position samples are contained inside of the W^{PAW}. The AEGC is evaluated over the available eye-gaze-position sample space.

$$AEGC = \frac{100}{M}\sum_{k=1}^{M} GAZE^{W^{PAFW}}(k) \quad \textbf{(7)}$$

Variable $GAZE^{W^{PAFW}}(k)$ equals to one in the case if the k^{th} eye-gaze-position sample is contained by the W^{PAW} and it equals zero otherwise. M is the number of all eye-gaze-position samples collected over the whole test video.

Eye fixations are the key validation parameter in the majority of today's eye-tracker-based systems. The HESA based RTPCS is evaluated through an eye fixation analysis with two different minimum duration periods: 100 ms. and 150 ms.

The *Average Eye Fixation Containment (*AEFC) is calculated as a percentage of the eye fixation samples contained within W^{PAW}.

$$AEFC = \frac{100}{N}\sum_{k=1}^{N} FIX^{W^{PAFW}}(k) \quad \textbf{(8)}$$

k is the instance of time when $W^{PAW}(k)$ window was constructed. Variable $FIX^{W^{PAFW}}(k)$ equals 1 or 0. It equals one in the case if the k^{th} eye fixation sample is inside of W^{PAW} and it equals zero otherwise. N is the number of corresponding eye fixation samples that the AEFC is measured over. Due to the fact that not all eye-gaze-position samples are the part of the eye fixations, N presents only samples belonging to the eye fixations.

Examples where eye fixations are contained and missed by the W^{PAW} are presented in Figure 6.

4.8 Perceptual Resolution Gain

The actual amount of bandwidth reduction and computational burden reduction when using the W^{PAW} depends on the two parameters: the size of the area which requires high quality coding (size of W^{PAW} for each frame) and visual degradation of the periphery. The *Average Perceptual Resolution Gain* (APRG) mathematically calculates the amount of additional compression achieved by a variable bit-rate RTPCS with the feedback loop delay.

$$APRG = \frac{M * H * W}{\sum_{t=1}^{M} \int_{0}^{W} \int_{0}^{H} S_t(x, y) dx dy} \quad \textbf{(9)}$$

$S_t(x, y)$ – is the eye sensitivity function represented by (2). An eye can see approximately one degree of the visual angle in highest-quality from the center of an eye fixation. Addressing the situation when the center of an eye fixation falls on the boundary of W^{PAW} one degree of visual angle is added to each dimension of the W^{PAW}. W and H are the width and the height of the visual image.

5 RESULTS

The experiment results evaluating the performance of the HESA based RTPCS are presented in Figure 7, Figure 8, and Figure 9.

5.1 Eye fixation containment

The AEFC performance of the HESA based RTPCS was evaluated for the case when

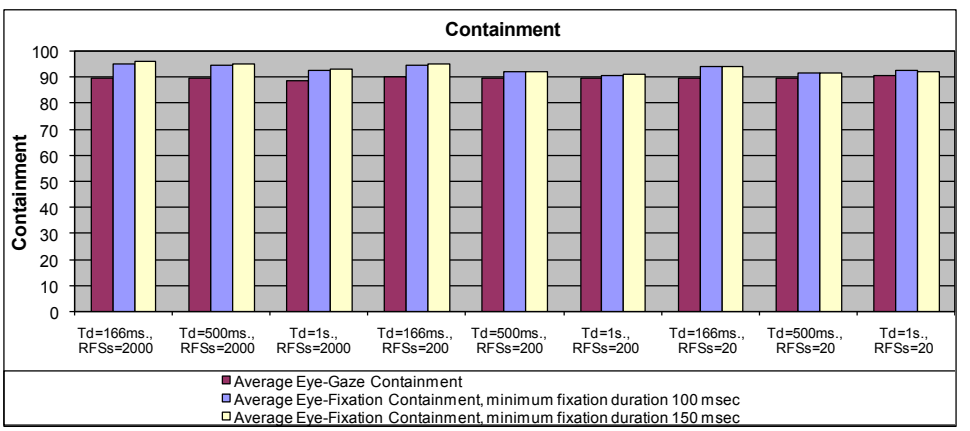


Figure 7. This figure presents the average eye-gaze containment versus average eye fixation containment considering eye fixation duration of 100 ms. and 150 ms. The X axis presents the feedback loop delay duration and RFSs value. The Y axis presents the containment values. The data points represent the average containment for the eye-gazes and two parameterizations of an eye fixation. AEGC and AEFC values are averaged between the subjects and test video clips.

the AEGC was approximately 90%.

The Figure 7 shows a side by side comparison of the AEGC and the AEFC calculation for two eye fixation parameterizations. On average the AEFC was always higher than the corresponding AEGC for all delay and the RFSs values. Standard deviation for the AEGC-AEFC values considering all subjects and video clips did not exceed 3. These results show that a RTPCS design based on the AEGC yields more conservative containment results than a system designed around pure eye fixation analysis. While the parameterization of the eye fixation detection mechanism changes the AEFC results, the AEGC generally provides a lower boundary for various parameterization choices.

An eye fixation detection requires a 100 – 150 ms. eye-gaze-position sample buffer, while pure eye-gaze-position analysis is virtually buffer free. This fact and the conservative nature of the eye-gaze containment make it a better choice for an RTPCS design.

5.2 Average Perceptual Resolution Gain

Figure 8 reports the APRG values for the test video set and the RFSs=2000 with the AEGC of 90%. The APRG values do not depend on the eye fixation parameterization.

The APRG varied from 1.95 for a delay of 3 ms. to 1.13 for a delay of 2 s. The delay of 500 ms. was the mark when the APRG decreased rapidly to 1.2.

Standard deviation of the APRG for different video clips fluctuated between 0.03 and

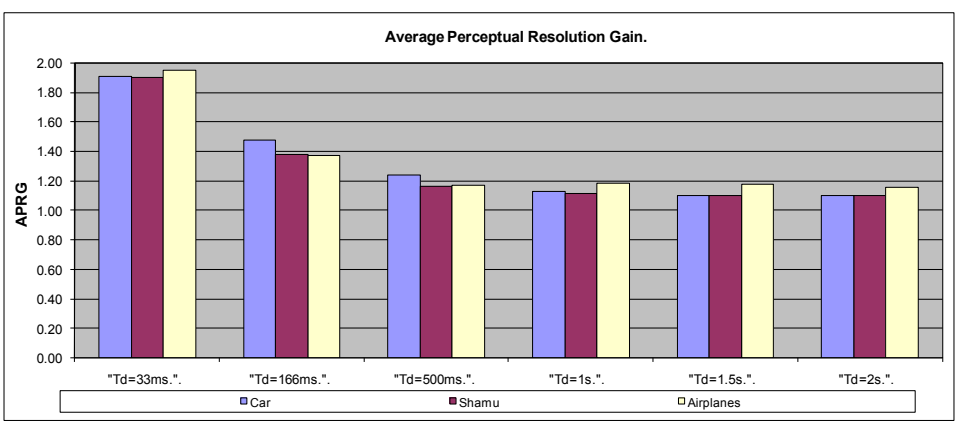


Figure 8. Average perceptual resolution gain achieved by the W^{PAW} constructed by the HESA model. The X axis presents the feedback loop delay duration in the system. The Y axis shows the APRG achieved by our RTPCS. The data points represent different video clips compressed by the system. The HESA model uses RFSs of 2000. The APRG values are averaged between the subjects.

0.11. Standard deviation of the APRG between subjects fluctuated from 0.02 to 0.45. This fact proves the perceptual compression is person and content dependant.

The “Airplanes” video yielded the best compression results for the high delays. The “Car” video had the best compression results for middle range delays of 166 - 500 ms. The “Shamu” video had the lowest compression results of the three videos. This is due to a “busy” scene with multiple perceptual activities involved. This type of content will provide less opportunity for perceptual compression, due to the more rapid eye movements.

The results show that the HESA based RTPCS provides a significant compression, but up to a certain delay value. We should also mention that the actual compression values or the reduction of computational burden will depend on the particular encoding scheme. A lot of modern codecs can encode the motionless part of the background with a very few bits reducing overall bit-rate that way; but in the case of a video where everything is moving (as in the “Shamu” video), modern codec will fail to reduce the bandwidth without visual quality loss. In a scenario such as this, W^{PAW} provides a specific region for high quality coding. This reasoning is supported by Figure 9, where APRG values for the videos with a moving background (“Shamu”, “Airplanes”) are almost the same as in the “Car“ video with a still background.

6 DISCUSSION AND FUTURE WORK

6.1 HESA input parameters

The HESA model has two main input parameters: the *Target Eye-Gaze Containment* (TGC) and the *Running Frame Eye-Speed samples* (RFSs).

Ideally the TGC ensures the amount of eye-gaze-positions to be contained inside of the W^{PAW}. The TGC goal is to “cut off” the Running Frame Eye-Speed values formed by the saccades. Usually the amount of saccades does not exceed 10% of the eye trace, thus 90% is a good starting value for the TGC. The results of our experiments show that TGC is a conservative parameter with the resulting AEGC values always higher than the TGC.

In practice it is beneficial to gradually reduce the TGC until the desirable value of the AEGC is achieved. In the design of our RTPCS, we used an algorithm where the TGC value was reduced one percent at a time until the desired AEGC value was reached. The success of such an approach can be judged from Figure 9. Depending on the delay value the TGC value had to be reduced to 80-47% before the AEGC of 90% was reached. The speed of the AEGC adjustment depended on the value of the RFSs, with lower RFSs values requiring fewer steps to bring the AEGC to the desired level. The difference range

was 12-25% for the RFSs of 20 and 18 - 44% for the RFSs of 2000. The lower RFSs values will adjust better to the rapid change of the content as it is indicated by the lower differences between the TGC and the AEGC numbers, but in case of a severe network jitter a W^{PAW} will be created with artificially small or large size, possibly making the compression artifacts more visible.

For the reference, Figure 9 shows the statistics of correlation between the TGC and the AEGC values recorded in our experiments. When the TGC equaled 90%, the AEGC tended to be very close to 100% with standard deviation of 1.2-10.1 between subjects and videos. When the TGC equaled 70%, the AEGC fluctuated between 71%-91% with stdev of 13-22. When the TGC equaled 60%, the AEGC fluctuated between 72%-86% with stdev of 17-24. From these results, it is possible to see that the correlation between the TGC and the actual AEGC is subject-, video- and visual-task dependant.

In our experiments, we found that the AEGC of 90% provides an acceptable compromise between high eye fixation containment and the additional compression received through the use of the W^{PAW}. Some other type of the multimedia might require a different selection of the AEGC number.

6.2 W^{PAW} accuracy

It is a valid question to ask how "bad" is it for the system which uses perceptual compression if an eye fixation is missed. The results might vary depending on how far the missed eye fixation is from the W^{PAW} boundary and what mapping of the eye sensitivity function presented by (1) is used for a particular choice of media source. If a viewer notices the "blurred" effect and is unable to see a specific detail on the picture, he

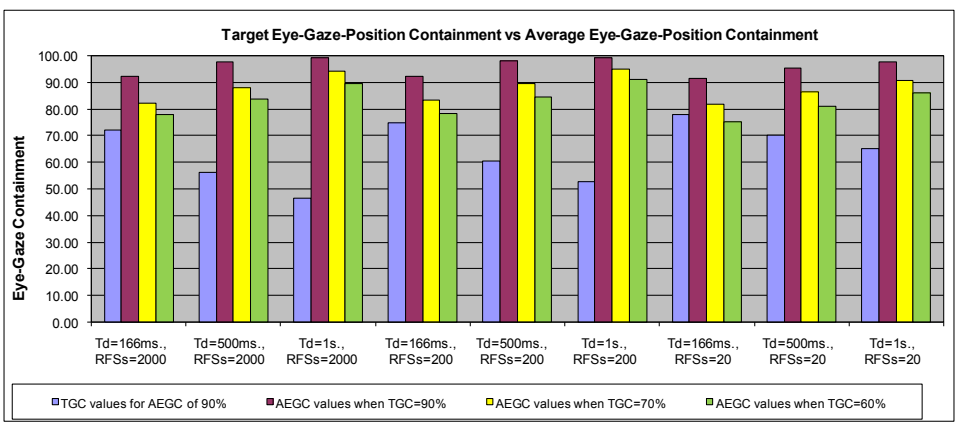


Figure 9. Correlation between the average eye-gaze-position containment and the target eye-gaze-position containment. The X axis shows the feedback loop delay and the RFSs value for each scenario. The Y shows the AEGC value. The TGC and the AEGC values are averaged out between subjects and videos.

or she can fixate their eyes on the point in question and the system will stabilize itself, placing the W^{PAW} on the spot under attention. The amount of time used during the stabilization will depend on the RFSs parameter and the feedback delay value.

6.3 Network challenges

It should be pointed out that there is an additional challenge for the detection of any eye movement type inside of an RTPCS. For example, there is an uneven delay variation in the multimedia compression mechanism because different types of video frames take various amounts of time to encode and process. When a network jitter exists, the raw eye-gaze position samples start arriving at different components of an RTPCS unevenly, thus making the interpretation of the eye movement types even more difficult. Additionally the transmission delay/lag (feedback loop delay) can be an order of magnitude larger than the duration of a basic eye movement (a saccade or an eye fixation). Thus by the time a current viewer's eye movement type is identified by RTPCS, that type of eye movement might be effectively over and useless for the prediction mechanism. Under the circumstances such as these, relying on the average eye-gaze containment instead of eye fixation containment is especially beneficial.

7 CONCLUSION

Perceptual compression will be critical in order to achieve the compression ratios needed in the emerging applications that require compression levels far beyond those available through the use of the classic compression methods. Perceptual multimedia compression might be especially viable in these scenarios: remote vehicle control and operation, remote surgery assistance, virtual reality teleporting or in applications where eye-gaze is used as an input or content evaluation tool.

In this paper, we have proposed an eye-gaze-position-based design and evaluation model of a Real Time Perceptual Compression System (RTPCS). Our results indicate that the eye-gaze-position containment (AEGC) is a more conservative evaluation metric than the eye fixation containment metric. Additionally, the AEGC does not require the eye-gaze-position buffering as it is required in the eye fixation case.

One of the most critical challenges in the design of an RTPCS design is the issue of the feedback loop delay. This issue has not been considered by the previous research. In this paper, we have addressed this issue through the concept of Perceptual Attention Window.

The important aspect of our research is that it is media independent. We have proposed a perceptual attention window as a virtual area superimposed on the rendering plane of any visual media. Once the window parameters are obtained, then the actual fovea-matched encoding can be performed in numerous media specific ways with various computational-effort/quality/rate trade-off efficiencies. Mathematical evaluation shows that the HESA based RTPCS is capable of compressing a multimedia source by up to 1.95 times.

8 ACKNOWLEDGEMENTS[1]

This work was supported in part by DARPA Research Grant F30602-99-1-0515.

9 REFERENCES

[1] Cormen T.H., Leiserson C.E., Rivest R.L. Introduction to Algorithms. MIT Press/McGraw-Hill, 1990.

[2] Irwin, D. E. Visual Memory Within and Across Fixations. In *Eye movements and Visual Cognition: Scene Preparation and Reading*, K. Rayner, Ed. Springer-Verlag, New-York, NY,1992, pp. 146-165. Springer Series in Neuropsychology.

[3] Komogortsev O., Khan J. Predictive Perceptual Compression for Real Time Video Communication. In *Proceedings of the 12th ACM International conference on Multimedia (ACM MM 04)*, New York, NY, 2004, 220-227.

[4] Murphy, H., Duchowski, A. T. Gaze-contingent level of detail rendering. In *EuroGraphics 2001*, EuroGraphics Association.

[5] Kortum, P., Geisler, W. S., *Implementation of a Foveated Image Coding System for Image Bandwidth Reduction.* In Proc. SPIE Vol. 2657, 1996, 350-360.

[6] Lee, S., Pattichis, M., Bovok, A. *Foveated Video Compression with Optimal Rate Control.* In IEEE Transaction of Image Processing, V. 10, n.7, July 2001, pp-977-992.

[7] Yarbus L., Eye Movements and Vision, Institute for Problems of Information Transmission Academy of Sciences of the USSR, Moscow (1967).

[8] Kuyel T., Geisler W., and Ghosh J., "Retinally reconstructed images (RRIs): digital images having a resolution match with the human eye," in *Proc. SPIE Vol. 3299*, 1998, 603-614.

[10] Duchowski, A. T. Acuity-Matching Resolution Degradation Through Wavelet Coefficient Scaling. In *IEEE Transactions on Image Processing 9 (8)*, 2000, 1437-1440.

[11] Duchowski, A. T. and McCormick, B. H. "Preattentive considerations for gaze-contingent image processing," in *Proc. SPIE Vol. 2411*, 1995, 128-139.

[12] Bergstrom P., *Eye Movement Controlled Image Coding*, PhD dissertation, Electrical Engineering, Linkoping University, Linkoping, Sweden, (2003).

[13] Geisler W. S. and Perry J. S. Real-time foveated multiresolution system for low-bandwidth video communication. In *Proc. SPIE Vol. 3299*, 1998, 294-305.

---

[1] We would like to thank the shepherd Dr. Milton Chen at Stanford and the anonymous reviewers of the "Predictive Perceptual Compression for Real Time Video Communication" paper for recommending that paper as a best student paper for the 12th ACM International Conference on Multimedia (ACM MM 04) conference. Their advice and guidance encouraged us to go further in our research and investigate additional aspects in this exciting area. New results that we present in this journal paper would not be possible without their wisdom.

[14] Daly S., Matthews K. and Ribas-Corbera J. As Plain as the Noise on Your Face: Adaptive Video Compression Using Face Detection and Visual Eccentricity Models. In *Journal of Electronic Imaging V. 10 (01)*, 2001, 30-46.

[15] Daly S. Engineering observations from spatiovelocity and spatiotemporal visual models. In *Proc. SPIE Vol. 3299*, 1998, 180-191.

[16] Tobii technology. User Manual. Tobii Eye Tracker ClearView analysis software. Copyright Tobii Technology AB, February 2006.

[17] Stelmach L. B., and Tam W. J. Processing image sequences based on eye movements. In *Proc. SPIE Vol. 2179*, 1994, 90-98.

[18] Babcock J.S., Pelz J.B., and Fairchild M.D. Eye tracking observers during color image evaluation tasks. In *Proc. SPIE Vol. 5007*, 2003, 218-230.

[19] Yanoff M. and Durker J. *Ophthalmology*. Mosby International Ltd., 1999.

[20] Shebilske, W. L., and Fisher, D. F. "Understanding Extended Discourse Through the Eyes: How and Why" In R. Groner, C. Menz, D. F. Fisher, and R. A. Monty (Eds.), Eye Movements and Psychological Functions: International Views. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983, pp. 303-314.

[21] Carmy, R. and Itti L. Casual Saliency Effects During Natural Vision. In *Proceedings of the symposium on Eye TrackingResearch & Applications 2006 (ETRA 06)*, (March 2006), 11-18.

[22] Peters, R. and Itti L. Computational mechanism for gaze direction in interactive visual environments. In *Proceedings of the symposium on Eye Tracking Research & Applications 2006 (ETRA 06)*, (March 2006), 27-32.

[23] Komogortsev O., Khan J. Perceptual Attention Focus Prediction for Multiple Viewers in Case of Multimedia Perceptual Compression with Feedback Delay. In *Proceedings of the symposium on Eye Tracking Research & Applications 2006 (ETRA 06)*, (March 2006), 101-108.

[24] Khan J., Komogortsev O. A Hybrid Scheme for Perceptual Object Window Design with Joint Scene Analysis and Eye-Gaze Tracking for Media Encoding based on Perceptual Attention. *In Journal of Electronic Imaging 15(02)*, April 2006, pp. 1-12.

[25] Duchowski A. T. *Eye Tracking Methodology: Theory and Practic.* Springer-Verlag, London, UK, (2003).

[26] ASL Laboratories. Eyenal (Eye-Analysis) software Manual Windows version for use with ASL Series 5000 and ETS-PC Eye Tracking Systems. Copyright 2001, by Applied Science Group, Inc.

[27] Komogortsev, O., Perceptual Test Video Set. At *http://www.cs.txstate.edu/~ok11/videosetpercept.htm*.

[28] Khan J., Yang S., Patel D., Komogortsev O., Oh W., Guo Z., Gu Q., Mail P., "Resource Adaptive Netcentric Systems on Active Network: a Self-Organizing Video Stream that Automorphs itself while in Transit Via a Quasi-Active Network", In Proceedings of the *Active Networks Conference and Exposition* (DANCE '2002), IEEE Computer Society Press, San Francisco, California, May 29-31, 2002, ISBN: 0-7695-1564-9, pp. 409-426.

[29] Irwin, D. E. "Visual Memory Within and Across Fixations". In K. Rayner (Ed.), *Eye movements and visual cognition: scene preparation and reading.* New-York, NY: Ed. Springer-Verlag, 1992, pp. 146-165. Springer Series in Neuropsychology.