Technical Report 2002-11-01
Internetworking and Media Communications Research Laboratories
Department of Computer Science, Kent State University
http://medianet.kent.edu/technicalreports.html

Dynamic Gaze Span Window based Foveation
for Perceptual Media Streaming

Javed I. Khan & Oleg Komogortsev

Abstract

The human vision offers a tremendous scope of data compression. Only about 2 degree in our about 140 degrees vision span has sharp vision. A fascinating body of research exists in vision and psychology geared towards the understanding of human visual perception system. The possibility of eye-tracking based perceptual compression has been anticipated for some time by many researchers. We have recently implemented one such system-- a live eye-gaze integrated media streaming system. It integrates a streaming server, a real-time live media transcoder and a live magnetic head-tracker integrated high-speed eye tracker. A unique challenge of this real time perceptual streaming is how to handle the fast nature of human eye-gaze interaction with relatively complex MPEG-2 rate transcoding scheme, and the control loop delay associated with streaming in the network. We have designed a live eye gaze interaction based dynamic foveation windowing scheme to address the challenge.

1 Introduction

The human vision offers a tremendous scope of perceptual data reduction. Only about 2 degree in our about 140 degrees vision span has sharp vision. There are two kinds of photoreceptor cells in a human eye playing the crucial sensory role for our vision: rods and cones. The rods are more sensitive to light and function primarily during night vision. Cones provide fine-grained spatial resolvability of the visual system. These photoreceptor cells make connection to the ganglion cells and their axons form the optical nerves. Photoreceptor cells are non-uniformly distributed with most concentration in the central part of the retina. Human acuity perception is related to their sampling and to their mapping to the ganglion cells. The diameter of the highest acuity circular region subtends only 2 degrees, the parafovea (zone of high density) extends to about 4 to 5 degrees, and acuity drops off sharply beyond. At 5 degrees it is only 50% [Irw92].

A fascinating body of research exists in vision and psychology geared towards the understanding of human visual perception system. This is extremely complex area and only part of it is believed to be understood. In this research we are exploring if any of the knowledge gained can be translated into direct engineering benefit. We particularly focus on the case of perceptual data reduction in projection media streaming and presentation. A whole new form of human computer interaction can be created by the eye-tracking based systems. Such eye-gaze based media streaming can be precursor to a new generation of high fidelity, and large aperture immersive virtual reality type systems- which takes human movement, gaze and attention into consideration.

A particularly important factor in such integrated perceptual encoding is the delay between the time an eye-gaze can be tracked and the time by which the coding response arrives at the screen. This delay is particularly significant in systems which involve network transmission. It is also substantial when large format media are to be perceptually transformed. In this report we particularly address how this delay can be managed by a novel dynamic windowing mechanism on the visual plane.

1.1 Related Work

Over the years many important issues related to foveation have been studied. A particularly active area is the study of various contrast sensitivity or spatial degradation models around the foveation center and their impact on the perceived loss of quality by subjects [DuMB98, KUGG98, LoMc00, Duch00b]. These studies observed potential for significant bandwidth reduction even up to 94.7%. Also various probable techniques for variable spatial resolution coding have been a topic of vigorous research. Examples include Wavelet-based Spatial Coding [Niu95, WLB01], Resolution Grid [KoGW96], Retinal Coding [KuGG98], etc. Several investigations studied CSF and coding techniques for videos in particular [KhYu96a, Yoor97, WSLR97, DuMB98, GWPJ98, Daly98, WaLB01, LePB01], though, none reported integration of live eye-tracker with video live coding. [DMKR99] studied facial video, and instead of eye-gaze used image analysis to detect focus. [GWPJ98] presented pyramid coding with pointing device to identify focus. [KhYu96a, KhYu96b] suggested mouse driven interaction for medical visualization. Computational complexity of video transcoding has also been deemed as a challenge and it confined earlier experiments to low bit-rate smaller formats. Recently [WaLB01] discussed a solution to the frame prediction problem common in relatively faster compressed DCT domain transcoding techniques. [LePB01] discussed how to optimally control the bit-rate for MPEG-4/ H.263 stream for foveated encoding. Their simulation used a set of given fixation point(s) and predicted about 8-52% saving for I pictures and about 68% for P pictures for 352x288 video sequences in wireless environment.

1.2 Our Approach

In this context, we have recently completed the implementation of a live foveation integrated media transmission scheme. It integrates a real-time live media transcoder with a live-eye tracker. The system intakes live perceptual information related to subjects eye position and head-movement via an eye-tracker and a magnetic head tracker respectively and correspondingly controls the spatio-temporal resolution of the presentation. The eye-tracker tracks the eye-gaze with respect to the human head. And the magnetic head tracker detects the movement of the head with respect to the scene plane and together they determine the eye-movement with respect to the presentation.

The **Percept Media Transcoder** (PMT) unit separates the perceptual sensors from media specific perceptual encoding. The PMT architecture has been designed so that multiple media types and media specific perceptual transcoding modules can be plugged into it without requiring the reorganization of the overall media distribution systems networking. That architecture enables one to use any standard-based media server and presentation system. For example, if remote rendered 3D graphics is passing through the PMT the graphics computation can be steered and optimize accordingly. We have incorporated a new full logic MPEG-2 high resolution region-based full-logic motion-vector inferencing (MVI) transcoder plugin into the PMT, which is fast but does not suffer from drift problem.

A critical consideration for such a real-time perceptual feedback based media transcoding scheme is the feedback delay between the position detection of the eye and displaying the encoded frame for the future eye movements. It is important to note, that a network not only imposes control loop delay but also the delay is dynamically varying. Also the effect cannot be ignored. As we will show it seems to be playing a dominant role in the actual determination of fovea region. A typical network delay ranges from 20ms to few seconds. Saccades can move the eye position more than 10-100 degrees in that time potentially wiping out the entire advantage of designing an accurate acuity window within the 2 degrees of foveation.

Consequently we pursue an approach that can operate with dynamically varying delay in detection of interaction in its design. Here, instead of relying only on the *acuity matching* model, we propose the integrated approach of *gaze proximity prediction and containment*. We determine a gaze proximity zone or a *foveation containment window*. Its target is to ensure that bulk of the eye-gazes remain within the window with a statistical guarantee. This approach roughly divides the problem into two parts. The first (and more well studies problem) is (i) how to associate the parafoveal degradation at the boundary of the containment zone with specific spatial resolution (quantization value, color) on the display based on the specific media type and modality. (ii) The interaction delay adds another challenge - how to predict and estimate the gaze containment so that the spatial resolution can be applied in the right place. Many of the previous research did address the first part. While almost no literature exists on the second. In this report we focus here. We will present a technique— as we will show, that more than 90% of the gazes can be contained within 20-25% window coverage area. The central consideration of this study is the feedback delay between the eye-tracking, coding, and presentation.

The report is organized in the following way. In next section we begin by describes briefly the human eye movement characteristics that played important role in this gaze containment algorithm. Then in section 2 we present the construction of the acuity model and the dynamic eye-movements tracking model, and the algorithm. Section- 4 then presents the containment performance from live experiment. In concept this virtual window can be applied to perceptually alter any visual media type. However, to help understanding the setup and architecture, in section-3 we briefly explain the MPEG-2 specific video transcoding and rate control used in this experiment.

2 Reflex Windowing

2.1 Human Visual Dynamics

The eye-containment system we propose is deeply related to the movement of human eye. Scientists have identified several intricate types of eye movements such as drift, saccades, fixation, smooth pursuit eye-movement, involuntary saccades. Among them, saccades and fixations play important role in the design of the proposed system.

Saccades: The eye moments that occur between two points of fixations (to be explained shortly) are called saccades. They are accomplished by eye movements of a single type – identical and simultaneous very rapid rotations of the eyes. Amplitude of the saccade usually doesn't exceed 20 degree. For angels less that 1 degree the duration of the saccade is 0.01-0.02 sec; for angles of 20 degrees it may reach 0.06-0.07 sec. The maximum velocity reached by the eye during a saccade of 20 degrees is 450degr/sec.

Fixations: Several types of eye movements also take place when the object of perception is stationary relative to the observer's head. Human's eye moves in three way during fixation: by small involuntary saccades, equal for the two eyes; by drift, slow, irregular movement of the optical axes in which however some degree of constancy of their position is retained; and by tremor, an oscillatory movements of the eyes of high frequency but low amplitude. The amplitude of the tremor is 20-40 seconds of the angle. Frequency of the tremor movements is 70-90 oscillations per second. Small, involuntary saccades appear if fixation on the object exceeds certain length – usually 0.3-0.5 sec. The duration of the drift usually last somewhere around 0.3 sec. During long fixation 97% of time is drift and only 3% small involuntary saccades [Yarb67].

2.2 Overview of the Scheme

The approach we take is first derive an parafoveal window based on acuity - eye sensitivity function, then add correction that takes into consideration the reflex eye-movement between the time the eye is tracked, and the perceptually encoded frame would appear in the display. We define three windows:

To address the eye acuity we define the $W_A(t)$ *acuity window* (AW). AW addresses bit distribution scheme matching HVS during eye fixation.

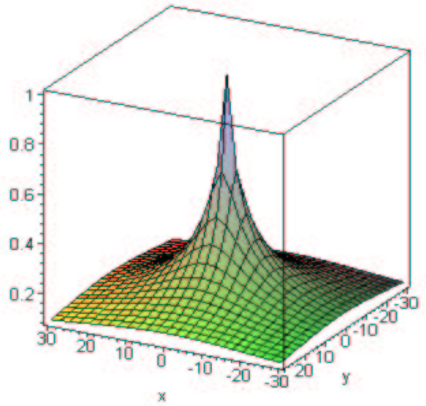

Fig. 1 Visual sensitivity function of the human eye.

To predict future eye movements we present $W_R(t)$ *reflex window* (RW). RW represents a container for saccadic eye-movements and it represents the area, where AWs should be placed in the future.

We present $W(t) = f(W_A(t), W_R(t))$ *visual window* (VW), as a combination of AWs and RW. VW provides imperceptible HVS enhanced video compression in real time encoding situation.

In the next three subsections we correspondingly describe the design on these three windows.

2.3 Visual Sensitivity - Acuity Model

We already mentioned the perception of acuity depends of the spatial distribution and mapping of cones, rods and ganglion cells, and the mapping of the visual fields across the visual context. A large number of functions have been suggested for *contrast sensitivity function* (CSF). Some of them are based on anatomical considerations and some are based of psychovisual empirical studies. For this experiment we have used equation 2.3 which has been modeled after the CSF function presented by [DMKR99]. This is based on entropy losses of the visual system. Also, this function the issues of cones, rods, ganglion cells distributions and based on two sets of data provided by [ViRo79] and [John87]. In our model, we will also adapt this function. Fig-1 shows the acuity distribution in the visual plane.

$$S(x, y) = \frac{1}{1 + k_{ECC} \cdot \theta_E(x, y)} \qquad \ldots\ldots(2.3)$$

Here S is the visual sensitivity as function of the video frame position (x,y) , k_{ECC} is a constant (in this model $k_{ECC} = 0.24$), and $\theta_E(x, y)$ is the eccentricity in visual angle. Within any lossy compression scheme the acuity quantity S have to mapped to the spatial loss or fidelity control functions of the scheme. The case of MPEG-2 will be described shortly.

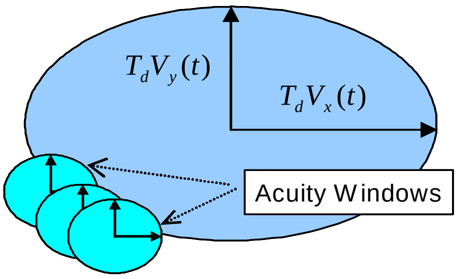

Fig. 2 Reflex window covered by a set of acuity windows

2.4 Reflex Window

Now we model the Reflex window. The objective of the reflex window is to contain the fixations by estimating the probable maximum possible eye velocity due to saccades. Given a set of past eye-positions, the reflex window predicts a zone where the eye will be at a certain point in future with target likelihood. We model the reflex window as an ellipse with focuses $\{T_d \cdot V_x(t), T_d \cdot V_y(t)\}$ where T_d is *prediction delay* (referred as FD in the graphs) and $V_x(t)$ and $V_y(t)$ are *containment assured velocity* (CAV).

2.5 Combining Acuity and Reflex Windows

Finally we model the combined window that can take into account both the acuity distribution as well as the eye motion. The eye-velocity determined the reflex ellipse. The eye is expected to be anywhere within this region. The acuity can further add boundary to this reflex window. We calculate the visual sensitivity as a function of eccentricity. We assume the subject's eye will be directed anywhere within RW with equal likelihood, the eccentricity is then measured from RW boundary and described as:

$$\theta_E(x,y) = \frac{180}{\pi}\arctan\left(\frac{\sqrt{\left(\frac{x-x_C}{x_R/y_R}\right)^2 + (y-y_C)^2} - y_R}{VD}\right) \qquad \ldots(2.4)$$

where x and y here are horizontal and vertical pixel position on the video frame, VD is the viewing distance in the units of pixel spacing. Quantities x_C and y_C are the coordinates of the center of the RW window, and $x_R = T_d \cdot V_x(t)$ and $y_R = T_d \cdot V_y(t)$ are the dimensions of the reflex window. Thus the prediction corrected sensitivity function is:

$$S(x,y) = \frac{1}{1 + k_{ECC} \cdot \frac{180}{\pi}\arctan\left(\frac{\sqrt{\left(\frac{x-x_C}{V_x(t)/V_y(t)}\right)^2 + (y-y_C)^2} - V_y(t)\cdot T_d}{VD}\right)} \qquad (2.5)$$

We can calculate visual sensitivity for each pixel using formulas 2.3 and 2.5. As a result we would have a perceptually encoded image with the give sensitivity resolution.

2.6 Dynamic Reflex Window Construction

Now we will state the velocity prediction method. Based, on the past positional variances we would like to estimate what should be the right velocity components to be used in equation-2.5a for a given prediction accuracy goal. We use the following k-median algorithm to determine this.

We are going to use a set of bins each for collecting velocity samples that fall in a particular velocity range. But first, we should say a couple of words about the speed of the sample collection process we use in our algorithm: In our model we estimate the velocity as an average on n past samples. Suppose there are n eye gazes during t-th frame. Each eye gaze S(t) has (x,y) position on the frame F(t) (position in units of pixels). The estimated horizontal and vertical components of the eye velocity are then estimated as:

$$\hat{V}_x(t) = \frac{1}{n}\sum_{i=1}^{n} | x(t-i-T_d) - x(t-i-1-T_d) | \qquad (4.5a)$$

$$\hat{V}_y(t) = \frac{1}{n}\sum_{i=1}^{n} | y(t-i-T_d) - y(t-i-1-T_d) | \qquad (4.5b)$$

We will call these samples *running average velocity* (RAV).

How fast do human eyes can move? Fig-3 shows a sample eye velocity measurement of a subject on various frames in degrees/frame. Here x-axis shows frames and y-axis shows the degrees/frame – how much/far eye moved during one frame.

Values $\hat{V}_x(t)$ and $\hat{V}_y(t)$ are originally calculated as pixel position. Given L (inches), is the distance between subject eyes and the display, H_{ph} (inches) is the horizontal image size, W_{ph} (inches) is the vertical image size, H (pixels) is the horizontal image size, and W (pixels) is the vertical image size we convert $\hat{V}_x(t)$ and $\hat{V}_y(t)$ to angular (degrees) values:

$$V_{x\_angular}(t) = \arctan\left(\frac{H_{ph} * \hat{V}_x(t)}{L * H}\right) \qquad \ldots\ldots(4.5c)$$

$$V_{y\_angular}(t) = \arctan\left(\frac{V_{ph} * \hat{V}_y(t)}{L * W}\right) \qquad \ldots\ldots(4.5d)$$

In real implementation the center of reflex window is placed on the last available eye-gaze. The equations for RW center would be: $x(t-n-T_d)$ and $y(t-n-T_d)$. Delayed eye gazes will be represented by: $x(t-i-T_d)$ and, where $0 \le i \le n$, and n is number of eye samples for the frame $F(t-T_d)$. Real eye gazes coordinates would be: $x(t-n)$ and $y(t-n)$, where n is number of eye samples for frame $F(t)$. As these velocity samples arrive they are collected in the bins. Let's consider two sets of bins. For each arriving eye-sample the x and y speed components are accounted separately. Let these sets be:

$X = \{x_0,..., x_i : \; x_k \geq 0; \; 0 \leq i \leq W\}$ and
$Y = \{y_0,..., y_j : \; y_k \geq 0; \; 0 \leq j \leq H\}$.

Sets X (or Y) have an associated velocity $\hat{V}_{i,x}(t)$ (or$\hat{V}_{j,y}(t)$) and an associated size count $C_{i,x}(t)$ (or $C_{j,y}(t)$). A sample belongs to an x_i if it has $\hat{V}_{i,x}(t) \leq \hat{V}_x(t) < \hat{V}_{i-1,x}(t)$ (and to y_i if it has $\hat{V}_{j,y}(t) \leq \hat{V}_y(t) < \hat{V}_{j-1,y}(t)$ for y component). Raw samples however are not used. They are smoothed as per equation-4.5 a & b. $\hat{V}_x(t)$ and $\hat{V}_y(t)$ are integer values; they are used as indexes for updating the counter values. The sizes of X and Y are chosen to be W and H respectively (dimensions of video image in pixels). With each incoming sample a counters in specific velocity bin is incremented.

$$c_{i,x} = c_{i,x} + 1 \quad \text{and} \qquad \text{……(4.5e)}$$
$$c_{i,y} = c_{i,y} + 1$$

We also include a history limit, where any samples older than "k" frame units (this value is referenced as velocity samples (VS) in the graphs) are discarded. This is accomplished by setting up a circular queue and count update scheme as following:

$$c_{i-k,x} = c_{i-k,x} - 1 \qquad \text{……(4.5f)}$$
$$c_{j-k,x} = c_{j-k,x} - 1$$

Let, ϖ to be *target containment factor*. ϖ corresponds to the amount of gazes to be contained in RW. $\varpi \in (0,..,1]$. Fro example $\varpi = 0.8$ would mean that we would to contain 80% of gases in RW. Then we can determine the maximum m for which:

$$\sum_{i=1}^{\max(m)} x_i \leq \varpi \sum_{i=1}^{W} x_i \qquad \text{……(4.5g)}$$

Then corresponding $V_x(t) = \hat{V}(t)_{m,x}$. Similarly we determine maximum n for which,

$$\sum_{i=1}^{\max(n)} y_i \leq \varpi \sum_{i=1}^{H} y_i \qquad \text{……(4.5h)}$$

Then the corresponding $V_y(t) = \hat{V}(t)_{n,y}$. $V_x(t)$ and $V_y(t)$ we will call *containment assured velocity* (CAV).

3 MPEG-2 Full Logic Transcoding

Once, the visual window is obtained it is applied in the target media. Unlike compressed domain transcoding schemes [YoSX99], we have implemented a full logic transcoder with *motion vector inference* (MVI) developed by [KhGu01], which is fast but does not suffer from drift problem like the compressed domain transcoders. Drift occurs due to the accumulation of reference error in the predictions of P frames. This is particularly problematic for region based encoding as it deliberately takes out bits from references. Our transcoder employs a full decoder and a MVI re-encoder. The MVI avoids the most computation intensive MV estimation process. Instead it reads and appropriately transforms the motion vector matrix from the incoming stream for a frame. Each transcoder transformation is designed as a pair-wise functions/algorithm {T^F(), and T^{MV}()}, where F_{out}=$T^F(F_{in})$, and MV^{out}=$T^{MV}(MV_{in})$. The MV^{out} matrix is used before the new prediction stage. Thus this scheme avoids any drift as well as the costly MV search. Also an enhanced TM-5 rate controller has been designed to allocate bits as per macroblock and frame types. The details of this transcoder transformation architecture and its rate control techniques are given in [KhGU01].

4 Experiment

4.1 Setup

We have implemented the system with integrated Applied Science Laboratories High speed Eye tracker Model 501. The eye position video capturing camera worked at the rate of 120 samples per second. For this experiment we defined fixation when the eye does not move more than 1 degree in 100msec.

A particularly challenging aspect of experimentation with perceptual system is the difficulty to model the subjective aspects of the human interaction. There is no agreed method. In this report we have taken an approach of carefully designing a set of objective parameters (*containment factor, goodness of containment, perceptual coverage*) as design target. Then we repeated the experiments on a large pool of carefully selected test videos, each offering various subjective challenges. We avoided presenting any

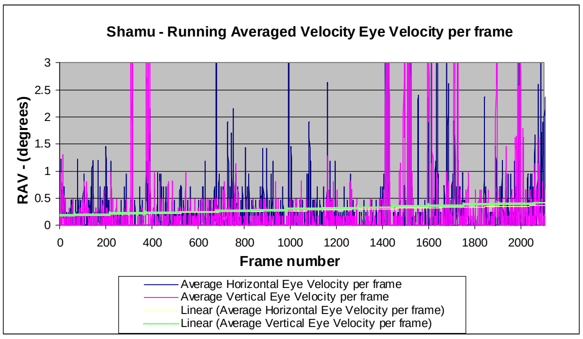

Fig. 3 Average angular eye speed for each frame (measured by linear algorithm described above)

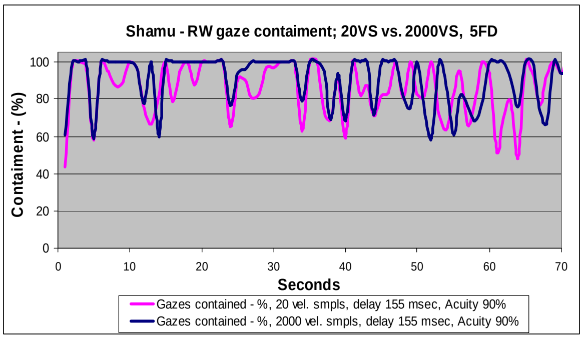

Fig. 5 Percentage of the eye gazes contained for 30 frames delay situation and two different schemes: 20 velocity samples and 2000 vel. smpl. are in consideration

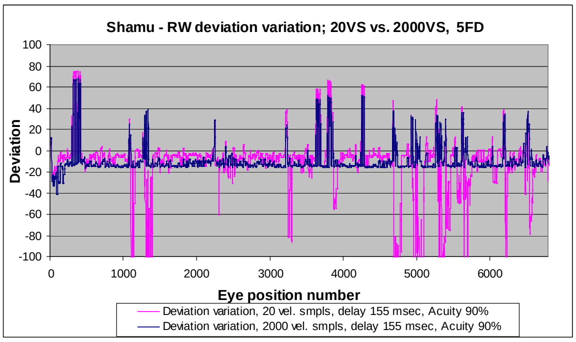

Fig. 6 RW deviation variation for two different schemes: 20 velocity samples and 2000 vel. smpl. are in consideration

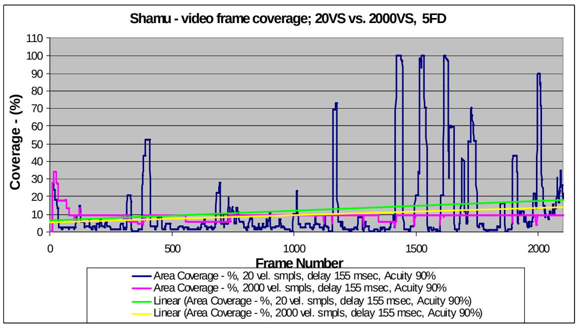

Fig. 7 Frame coverage by RW window for two different schemes: 20 velocity samples and 2000 vel. smpl. are considered

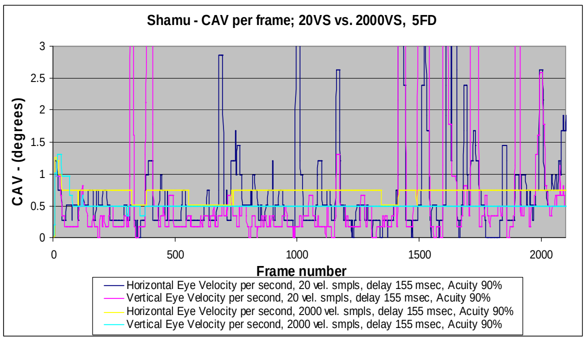

Fig. 8 Predicted containment eye velocity.

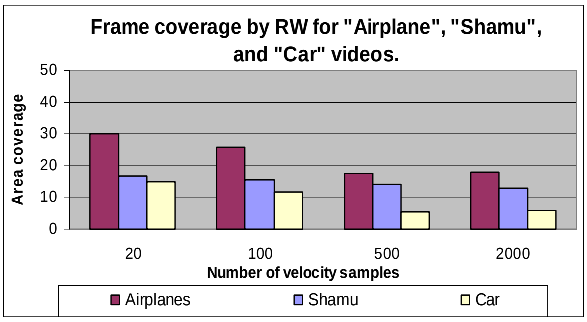

Fig. 9 Video frame coverage for three videos. System delay is 155ms. Axis "x" shows how many velocity samples where taking into consideration by dynamic RW building algorithm.

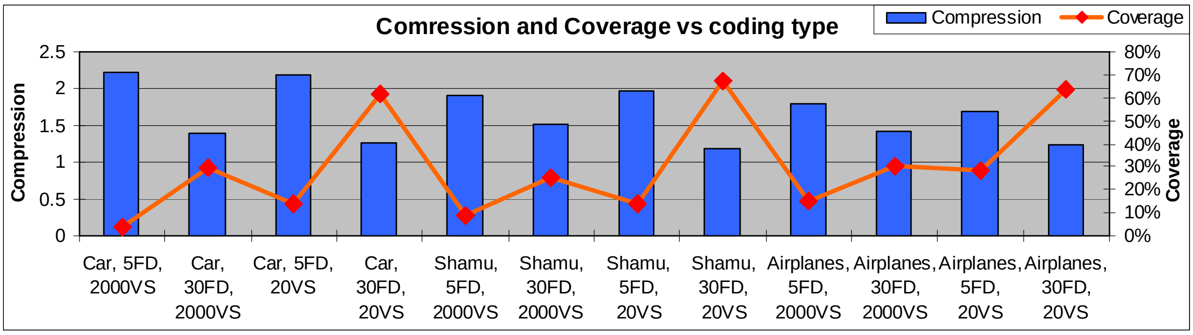

Fig. 10 Compression estimation and coverage results for different videos, system delays and dynamic velocity samples.

gross statistical average. Rather, we first describe a median performance. Then we present case-by-case analyses for the sequences.

All are videos were 720x480 and were captured with Sony TRV20 digital camera at high resolution with more than 500 lines at a frame rate of 30 frames per second. Number of frames per GOP is 15. Number of "B" frames between any give two "P" frames is two. The video was projected on the wall of the dark room. The projected physical dimensions of the image are width 60 inches, height 50 inches, and the distance between subject eyes and the surface of the screen is about 100-120 inches.

4.2 Gaze Containment

The first experiment we conducted is to see how effectively we were able to contain the gazes within the foveation window designed. We defined the quantity *gaze containment* as the fraction of gazes successfully contained within the window:

$$\xi = \frac{|S^{w}(t)|}{|S(t)|} \qquad \ldots\ldots(6.1a)$$

Where, W(t) is the entire sample set and $S^{W}(t) \subseteq S(t)$ be the sample subset contained within the window W(t). Fig-5 plots the results for target *containment factor* ω=0.9 (90%). As evident in most of the times the algorithm was able to contain 100% of the gazes. We studied the containment results for various *prediction delays* (FD) and *velocity sample memories* (VS). Fig-5 plots then per frame basis. These cases show excellent 90% average containment of the reflex window estimation, with rarely dropping to less than 60%.

We defined a quantity called *deviation*. For each sample s, we draw a line from it to the window center $m_W(t)$. Let us compute k as the intersection with the boundary of the window. The signed distance d(s,k) between k and s(t) is the deviation (Fig-4).

$$\delta = \begin{cases} d(s,k_s); \text{ if s(t) is outside } W_R(t) \\ -d(s,k_s); \text{ if s(t) is inside } W_R(t) \end{cases} \qquad \ldots\ldots(6.1b)$$

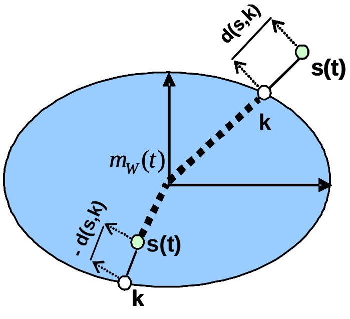

Fig. 4 Example for deviation calculation

If the number is negative (positive), than the gaze sample is inside (outside) the window. The magnitude identifies how far it is from the border. Fig-6 now plots the deviation for all samples for the same cases. As evident the deviation remained very close to the zero line. The window remained within 10% of the sample boundary. The deviation plot provides the optimality of the solutions; any smaller predicted window could have resulted in large number of misses.

4.3 Window Coverage Efficiency

Another important design goal is to reduce unnecessary containment as there will not be any perceptual redundancy to extract with large visual windows. We have defined a second performance parameter called "*perceptual coverage*" for obtaining coverage efficiency. If F(t) is the size of the total viewing frame, and W(t) is the predicted window, then the perceptual coverage is given by (delta for area or volume):

$$\chi(t) = \frac{|\Delta(W(t) \cup F(t))|}{|\Delta(F(t))|} \qquad \ldots\ldots.(6.3a)$$

Fig-7 shows the perceptual coverage for the cases. A system operating with about 155 ms delay would require only 10% of the frame to be encoded with high resolution. As indicated, a key factor that determined the size of the reflex window in our algorithm is the *containment assured velocity* (CAV). It will be interesting to see the CAV velocities estimated by the algorithm. Fig-8 plots the recorded CAVs for these cases in units of angular eye velocity. While, the smooth pursuit velocity was recorded in the range of .7 degree/sec, there were occasional fluctuations while the velocity shot up to 1 degree/sec and in some cases beyond 3 degrees/seconds. The larger fluctuations are believed to be caused by the eye-blinks. Two finer observations can be made. In both FD=5 and 30 delays, we can see that CAV looks like an averaged quantity on the first graph for 2000VS, but for 20VS it fluctuates a significantly.

4.4 Impact of Content Complexity

For insight into the impact of subjective complexity, here we present three representative case analyses-- one apparently simpler and one harder than the original

Shamu clip. Below are their rough subjective complexity descriptions:

- **Car in Parking Lot:** is a video of the moving car on the parking lot taken from a security camera view point in the university parking lot. The visible size of the car is approximately one fifth of the screen. Car moves slowly, letting subject to develop smooth pursuit movement (our assumption). Nothing on the background of this video distracts subject attention. Video duration is 1min 10sec.
- **Shamu:** This video captures an evening performance of Shamu at the Sea World, Ohio, under a tracking spotlight. It has several moving objects: shamu, trainer, and crowd. Each of them moving at the different speed 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 variety of eye-moments: fixations, saccades, pursuit. Such environment suits the goal of challenging our algorithm for different eye movements. Video duration is 2 mins.
- **Airplanes:** This video depicts formation flying of Ultrasonic planes – performed by Blue Angels over Lake Erie, rapidly changing their flying speeds. Number of planes varies from one to five for duration of the video. Also, the camera action involves rapid zoom and panning. This video provides a challenge for our algorithm to build a compact window to contain rapid eye-movements of the saccades and pursuit. Sometimes camera could not focus very well on the planes while capturing this video and subject has to search for the object. This aspect brings additional complication to the general pattern of eye movements for this video. This video duration is 1: min and 9 sec.

The originals and encoded versions of these videos are available at [KoKh02] for direct appreciation of the complexity. Fig-9 plots the "area coverage" obtained by the algorithm. We can see that, with this algorithm the reflex window was tightest on the "Car" video due to the smooth moving nature of the object inside the video. At VS=20 samples, the window on the average was about only 15%. The performance on "Shamu" video with more rapid and complicated object movements was next best 17%. "Airplanes" as expected gives worst performance of 30% due to the rapid ultrasonic airplanes movements inside the video and focusing problems what subject experienced while looking at the videos. Interestingly, the number of samples considered seems to have effect on the reflex-widow coverage. Longer memory with larger velocity samples seems to have considerable effect in improving the coverage efficiency to 7%, 13% and 17% for the three videos respectively.

5 Conclusions & Current Work

In this report we have explored the possibility of a human computer interaction based perceptual media streaming via a live system. The results from this system suggest some interesting discourse. Currently mainstream research has heavy concentration on the accurate design of CSF. Our experiment suggests that CSF will play lesser role in a movie. The delay in control loop (the delay in network, delay in media encoding, or even the delay within the eye-tracker) will create several times larger window of uncertainty. Thus, a simpler approximation of CSF will probably be as good as a detailed one overall.

In concept the virtual window we designed can be applied to any visual media type. If the visual window is successful in containing all the fixations, then the outer regions can be coded with very little bits without perceivable loss of quality. Once the window is obtained then fovea matched encoding can be performed in numerous media specific ways with various computational-effort/quality/rate trade-off efficiencies. Therefore, within the scope of this report, we refrained from media quality/rate discussion[1] rather focused on how the success of this system is in keeping the gazes contained within the visual window. However, the *intrinsic compressibility* of a presentation can be determined by estimating the sensitivity distribution over the video plane. Fig-10 plots it against the various coverage levels, and prediction delays for the test video set and suggest about 100-200% compressibility. The work is currently being funded by DARPA Research Grant F30602-99-1-0515.

6 References:

[Daly98] Daly, Scott J., "Engineering observations from spatiovelocity and spatiotemporal visual models" in *Human Vision and Electronic Imaging III*, July *1998, SPIE.*

[DMKR99] Daly, Scott J.; Matthews, Kristine E.; Ribas-Corbera, Jordi, "Visual eccentricity models in face-based video compression" in *Human Vision and Electronic Imaging IV*, May *1995, SPIE.*

[Duch00b] Duchowski, A.T., "Acuity-Matching Resolution Degradation Through Wavelet Coefficient Scaling. IEEE Transactions on Image Processing 9, 8. August 2000.

[1] Detail for MPEG-2 MVI transcoding and quality idea can be obtained from [KoKh02].

[DuMB98] Duchowski, A.T., McCormick, Bruce H., “Gaze-contingent video resolution degradation” in Human *Vision and Electronic Imaging III*, July 1998, SPIE.

[GWPJ98] Geisler, Wilson S.; Perry, Jeffrey S.; “Real-time foveated multiresolution system for low-bandwidth video communication**”** in *Human Vision and Electronic Imaging III, July 1998, SPIE.*

[Irw92] 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.

[ISO96] Information Technology- Generic Coding of Moving Pictures and Associated Audio Information: Video,ISO/IEC International Standard 13818-2, June 1996

[John87] A. Johnston, “Spatial scaling of central and peripheral contrast sensitivity functions”, *JOSA A V.4 #8*, 1987.

[KCKH95] Kim, Man-Bae; Cho, Yong-Duk; Kim, Dong-Kook; Ha, Nam-Kyu; “Compression of medical images with regions of interest (ROIs)” in *Visual Communications and Image Processing '95, April* 1995, SPIE.

[KhGu01] Javed I. Khan, Q. Gu, Network Aware Symbiotic Video Transcoding for Instream Rate Adaptation on Interactive Transport Control, IEEE International Symposium on Network Computing and Applications, IEEE NCA’ 2001, October 8-10, 2001, Cambridge, MA, pp.201-213.

[KhYu96a] Khan Javed I. & D. Yun, Multi Resolution Perceptual Encoding for Interactive Image Sharing in Remote Tele-Diagnostics, Manufacturing Agility and Hybrid Automation -I, Proceedings of the International Conference on Human Aspects of Advanced Manufacturing: Agility & Hybrid Automation, HAAMAHA'96, Maui, Hawaii, August 1996, pp183-187.

[KhYu96b] Khan Javed I. & D. Yun, Perceptual Focus Driven Image Transmission for Tele-Diagnostics, Proceedings of the International Conference on Computer Assisted Radiology, CAR'96, June 1996, pp579-584.

[KoGW96] Kortum, Philip; Geisler, Wilson S., “Implementation of a foveated image coding system for image bandwidth reduction” in *Human Vision and Electronic Imaging*, April *1996, SPIE.*

[KoKh02] Oleg Komogortsev and Javed I. Khan, Encoded Test Video Set from Dynamic Reflex Windowing, Technical Report 2002-06-01, Internetworking and Media Communications Research Laboratories, Department of Computer Science, Kent State University, http://medianet.kent.edu/technicalreports.html., June 2002.

[KuGG98] Kuyel, Turker; Geisler, Wilson S.; Ghosh, Joydeep, “Retinally reconstructed images (RRIs): digital images having a resolution match with the human eye” in *Human Vision and Electronic Imaging III*, July *1998, SPIE.*

[LePB01] S. Lee, M. Pattichis, A. Bovok, Foveated Video Compression with Optimal Rate Control, IEEE Transaction of Image Processing, V. 10, n.7, July 2001, pp-977-992.

[LoMc00] Lester C. Loschky; George W. McConkie, “User performance with gaze contingent multiresolutional displays” in *Eye tracking research & applications symposium*, November, 2000.

[Niu95] E.L. Niu, “Gaze-based video compression using wavelets”. University of Illinois at Urbana-Champaign. The Graduate College. August 1995.

[ViRo79] V. Virsu, J. Rovamo, “Visual resolution, contrast sensitivity, and the cortical magnification factor” in *Experimental Brain Research V. 37*, 1979.

[WaBo01] Z. Wang, and A. C. Bovik, “Embedded foveation image coding”, IEEE Trans. Image Proc., Oct. 2001

[WLB01] Z. Wang, Ligang Lu, and Alan C. Bovik, “Rate scalable video coding using a foveation-based human visual system model”, ICASSP 2001.

[WSLR97] Westen, Stefan J.; Lagendijk, Reginald L.; Biemond, Jan, “Spatio-temporal model of human vision for digital video compression” in Human Vision and Electronic Imaging II, June 1997, SPIE.

[Yarb67] L. Yarbus “Eye Movements and Vision” Institute for Problems of Information Transmission Academy of Sciences of the USSR, Moscow 1967.

[Yoor97] Seung Chul Yoon; Krishna Ratakonda; Narendra Ahuja, “Region-Based Video Coding Using A Multiscale Image Segmentation” in International Conference on Image Processing (ICIP '97), 1997.

[YoSX99] Youn, J, M.T. Sun, and J. Xin, "Video Transcoder Architectures for Bit Rate Scaling of H.263 Bit Streams," ACM Multimedia 1999’, Nov., 1999. pp243-250.