To appear in the Journal of Electronic Imaging, 2009

Gaze-Contingent Video Compression with Targeted Gaze Containment Performance

OLEG V. KOMOGORTSEV

Texas State University-San Marcos

Department of Computer Science

ok11@txstate.edu

This paper presents a delay compensation algorithm for a *Gaze-contingent video Compression System* (GCS) with a robust *Targeted Gaze Containment* (TGC) performance. The TGC parameter allows varying compression levels of a gaze-contingent video stream by controlling its perceptual quality. The delay compensation model presented in this paper is based on the Kalman filter framework that models human visual system with eye position and velocity data. The model predicts future eye position and constructs a high quality coded *Region of Interest* (ROI) designed to contain a targeted amount of gaze samples while reducing perceptual quality in the periphery of that region. Several model parameterization schemes were tested with 21 subjects and the delay range of 0.02-2 s and targeted gaze containment of 60-90%. The results indicate that the model was able to achieve targeted gaze containment levels with compression of 1.4-2.3 times for TGC=90% and compression of 1.8-2.5 for TGC=60%. Lowest compression values were recorded for high delays while highest compression values were reported during small delays.

Keywords: video compression, Kalman filter, eye movement prediction.

1 INTRODUCTION

Image compression has been a fruitful research area for many years. Modern methods use different techniques such as run-length encoding, entropy coding, chroma subsampling, transform coding, and motion compensation to achieve high ratios of compression in image and video. But the knowledge of the limitations of the *human visual system* (HVS) gives us more powerful means of improving already existing methods. The diameter of the eye's highest visual acuity, the fovea, extends only to 2 degrees. The parafovea, the next highest acuity zone, extends to about 4 to 5 degrees, and sharpness of vision drops off quickly beyond that point [1]. An HVS aware gaze-contingent compression offers tremendous potential for video bandwidth reduction [2-8]. In addition to bandwidth reduction, researchers report a drop of the computational burden for synthetic video and 3D imagery [4, 9]. As a result, gaze-contingent compression is quite a promising technique that can satisfy the needs of most recent bandwidth-hungry applications such as high-definition television, flight simulators, environment teleportation, virtual reality, telemedicine, remote vehicle operation, teleconferencing, etc.

The main idea behind gaze-contingent compression methods is to increase the resolution within the area of gaze point and to reduce the image quality in the periphery according to an acuity degradation function. The acuity-based peripheral degradation is possible due to characteristics of the light sensors and their distribution within the eye [10]. In a practical case of gaze-contingent compression systems, peripheral degradation has to be properly mapped to a specific method of compression (codec) to be perceptually lossless [11-13].

The information about gaze location is obtained by a device called an eye tracker [10]. The cost and accuracy of such devices have been improved recently. Still, many unsolved issues remain when eye trackers are employed in the realm of gaze-contingent compression. For example, point based peripheral degradation can work in a simple scenario in which an eye-tracker and a display unit are connected to the same computer. However, in situations that involve compression and transmission of the video from one point of the globe to another, the delay between gaze sensing and display can be quite significant. The delay effects would be noticeable when gaze lands on a not yet updated part of the image, thereby subjecting the viewer to the compression artifacts [14]. Therefore, if the future gaze location is predicted, a high quality ROI can be placed to "contain" future gaze coordinates, thus negating delay effects. It is obvious that the larger the amount of gaze samples are contained within an ROI, the higher perceptual quality the video will have. Therefore, it is important to have a parameter, such as Targeted Gaze Containment, that allows containing requested amounts of gaze samples within a ROI.

The idea of the targeted gaze containment was first investigated by Komogortsev and Khan [15]. The researchers created a *Histogram Eye-Speed Analysis* (HESA) model that allowed the ROI to contain a targeted amount of gaze samples given the value of the feedback delay. With empirically selected target gaze containment parameter, this method was able to contain approximately 90% of the gaze samples and provided estimated compression of up to 2 times in case of 33 ms and 1.4 times for 166 ms delay. Low compression factors of 1.1-1.2 were reported for the 1-2 s delay range, indicating a point of saturation for the HESA model. An additional weakness of the HESA model was the large difference between the TGC and the actual gaze-containment, sometimes reaching the level of 47%. The goal of this paper was to develop a model that would have more reliable performance in terms of the TGC and the actual gaze containment while providing higher compression values.

A new model proposed in this paper constructs ROI by a *Two State Kalman Filter* (TSKF), where the TSKF predicts future eye position by employing a Kalman filter given the value of the feedback delay. A Kalman filter is selected as a classical predictive framework that estimates the state of a dynamic system from a series of incomplete and noisy measurements. Parameters provided by the TSKF model constructs the ROI, assuming the highest resolution coding inside of this region and an HVS-based resolution degradation in the periphery. The engineering challenge related to the performance of any real-time RGC is to produce an optimum ROI where the targeted amount of gaze samples can be contained within a minimum visual space. Such a challenge is especially pertinent in case of a dynamic content where gaze behavior changes frequently. Addressing the challenge mentioned above, different TSKF parameterization approaches are presented in this paper.

2 RELATED WORK

A good summary of current research in the gaze-contingent display and compression fields can be found in [16-19]. A significant amount of gaze-contingent research was focused around spatial and temporal visual acuity degradation models and the implementation of those models for specific encoding schemes [11-13, 20, 21].

Itti and colleagues have devised several content analysis methods to detect image regions that might attract viewers' attention [22-24]. Those studies were directed to indentify the factors attracting visual attention in top down and bottom up scenarios of attention deployment. The attention maps called the saliency maps produced in this research can be employed in gaze-contingent compression systems by pre-encoding parts of the image with highest probability of attention using high resolution and reducing the resolution in the remaining part of the image.

The early research on the high quality coded ROI placement was done in the works of Duchowski and colleagues: [4, 20, 25, 26]. The main objective of that research was to place high quality coded regions on the part of the image that might be potentially interesting to the viewer and provide a way to encode the periphery which has lower quality with a method that prevents the detection of gaze-contingent compression artifacts. The ROI placement was either manually selected or calculated from the past gaze data recorded from several subjects' viewings - there was no real-time feedback between the gaze stream and the location/size of the ROI. The ROI radius was fixed to be 5°, based on the fact that acuity at 5° of eccentricity is roughly 50% of the acuity at the fovea [25].

An important study was conducted by Loschky and Wolverton that investigated the dependency between the image update delay in a gaze-contingent display and the detection ability of the compression artifacts [14]. The researchers have placed a fixed size ROI on top of the subjects' eye fixation after a pre-set interval and recorded the rate of detection of the peripheral "blurriness" by participants. A conclusion was made that if the image was updated within 60 ms after the beginning of a fixation, the gaze-contingent display provided perceptually lossless compression.

Several ROI placement strategies, with a direct feedback between current gaze data and the ROI placement/size, were investigated by Komogortsev and Khan: [7, 8, 15, 27, 28]. The main objective of that research was to place an ROI on the future gaze location to compensate for the sensor/transmission delay effects present in a gaze-contingent compression system. Methods for ROI calculation were broken into three categories: 1) ROI size/placement was determined only by current gaze stream and the delay value [7, 8, 15, 29]; 2) ROI size/placement was determined by hybrid methods that employed real-time extraction of object information from an MPEG-2 video stream and current gaze data with a delay value [27]; and 3) ROI size/placement was determined by the aggregation of several concurrent gaze data streams from multiple viewers [28]. The categories presented above can be also divided into two groups:

a) best effort – maximum amount of gaze samples are contained within the smallest ROI [8].

b) targeted performance - ROI size/placement attempted to contain targeted amount of gaze samples within the ROI given the value of the delay [7, 15, 27, 28]. The method presented in this paper belongs to the second category with an objective of creating an algorithm that will give better control over actual gaze containment, given the value of *Target Gaze Containment* (TGC), while providing higher compression levels.

3 HUMAN VISUAL SYSTEM AND GAZE-CONTINGENT COMPRESSION

The *human visual system* (HVS) exhibits a variety of eye movements: fixations, saccades, smooth pursuit, optokinetic reflex, vestibulo-ocular reflex, and vergence [30]. This paper concentrates on the first three. With great simplifications, their roles can be described as follows: fixation – eye movement that keeps gaze stable in regard to a stationary target, providing visual pictures with highest acuity; saccade – very rapid eye rotation moving the eye from one fixation point to another; and pursuit stabilizes the retina in regard to a moving object of interest [10]. Highest acuity vision occurs during fixations and pursuits. The vision is suppressed during saccades.

The HVS-based gaze-contingent compression approach adopted in this paper comes from the work of Daly et al. [5] and employs a visual sensitivity function that is related to the contrast sensitivity of the human eye:

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

Here, S is the eye visual sensitivity as a function of the image position (x,y), ECC is a constant (in this work ECC=0.24), and $\theta_E(x, y)$ is the eccentricity in the visual angle. Ideally such sensitivity function should provide perceptually lossless compression in a gaze-contingent system but the actual results will depend on the mapping between such a function and the actual compression algorithm [5]. The diagram of the eye sensitivity function is presented by Figure 1.

4 ROI DESIGN & APPLICATION

4.1 Delay Effects

The delay is defined as a period of time between the instance the eye position is detected by a sensor and the moment when the compressed image is displayed. The delay should be taken into consideration because future fixations/pursuits should fall within the highest quality region of the image, ensuring that gaze-contingent compression remains perceptually lossless. It is noteworthy that the properties of video transmission might change over time, thus increasing or decreasing the length of the delay. A network delay can be as high as a few seconds. As a result of the delay, saccades can place gaze on the low quality coded part of the image. Therefore, a real-time gaze-contingent system must have a prediction algorithm that allows placing high quality coded ROI on top of the future fixation/pursuit movement to compensate for the delay effects.

4.2 ROI Construction

The purpose of this study is to compensate for the delay effects by calculating the size and the placement of the high quality coded ROI and incorporate this ROI into the gaze-contingent compression model through visual sensitivity function. The objective of the ROI is to contain a targeted amount of gaze samples that belong to eye fixations and pursuits within the smallest possible area.

We move our eyes using saccades. The acceleration, rotation, and deceleration involved in ballistic saccades are guided by the muscle dynamics of the eye and demonstrate stable behavior. The latency, vector, direction of the gaze, and the eye-fixation duration have been found to be highly dependent on the content of the media presented in addition to being unpredictable. Therefore, we model the ROI as an ellipse, allowing the gaze to take any direction from the center. Each gaze contingent compression model has to define the center ($x_{cen}^{ROI}(k)$, $y_{cen}^{ROI}(k)$) and the length of the major, minor axis of the ROI ellipse ($x_{dim}^{ROI}(k)$, $y_{dim}^{ROI}(k)$). Figure 2 presents an example.

The *Two State Kalman Filter* (TSKF) model presented in this paper places the ROI center on top of the gaze position. The accuracy of such prediction and the amount of gaze samples previously contained by the ROI define its dimensions. Section 8 presents details.

4.3 ROI & Gaze-Contingent Compression

The logical structure represented by the ROI can be applied to bandwidth/computational burden reduction cases by reducing peripheral image characteristics though visual sensitivity function. To achieve this goal, a new visual sensitivity function incorporating the ROI concept needs to be computed. Assuming that the coordinates of the ROI center, the ROI dimensions, and the delay value T_d are provided, Equation (1) can be transformed into the form:

$$S_t(x_{pix}, y_{pix}) = \begin{cases} 1, \quad when \quad In\_Ellipse(t) < 0 \\ 1/\left(1 + ECC\dfrac{180}{\pi}tan^{-1}\left(\dfrac{In\_Ellipse(t)}{VD}\right)\right), \quad otherwise \end{cases} \tag{2}$$

where $In\_Ellipse(t) = \sqrt{\left(\frac{V_y(t)\left(x_{pix} - x_{cen}^{ROI}(t)\right)}{V_x(t)}\right)^2 + \left(y_{pix} - y_{cen}^{ROI}(t)\right)^2} - y_{dim}^{ROI}(t)$. x_{pix}, y_{pix} are coordinates of every pixel of the image presented. $x_{cen}^{ROI}(t)$, $y_{cen}^{ROI}(t)$ are the ROI center's coordinates and $x_{cen}^{ROI}(t)$, $y_{cen}^{ROI}(t)$ are the ROI's dimensions at the time instance "t". VD is the distance between the viewer's eyes and the screen surface. Note that all distances have to be converted to the pixel distances for this equation to be true.

The peak point presented by the Equation (1) and depicted by Figure 1 becomes the ellipse of the ROI depicted by Figure 2, therefore the role of $In\_Ellipse(t) < 0$ expression of the Equation (2) is to provide a check if a pixel with coordinates x_{pix}, y_{pix} is inside of the ROI ellipse or not. Every pixel inside of the ROI has to have the original visual quality with sensitivity value of 1. The slope of the visual sensitivity degradation is defined by the Equation (1) where eccentricity $\theta_E(x, y)$ is calculated to take into the account the ROI dimensions and the distance to the screen. Figure 3 provides a graphical representation of the Equation (2).

5 TWO STATE KALMAN FILTER MODEL

5.1 Basics of Kalman Filtering

The Kalman filter is a recursive estimator that computes a future estimate of the dynamic system state from a series of incomplete and noisy measurements. A Kalman Filter minimizes the error between the estimation of the system's state and the actual system's state. Only the estimated state from the previous time step and the new measurements are needed to compute the new state estimate. Many real dynamic systems do not exactly fit this model; however, because the Kalman filter is designed to operate in the presence of noise, an approximate fit is often adequate for the filter to be quite useful [31].

The Kalman filter addresses the problem of trying to estimate the state $x \in \Re^n$ of a discrete-time controlled process that is governed by the linear stochastic difference equation [31]:

$$x_{k+1} = A_{k+1}x_k + B_{k+1}u_{k+1} + w_{k+1} \tag{3}$$

with the measurement

$$z_k = H_k x_k + v_k \tag{4}$$

The n-by-n state transition matrix A_{k+1} relates the state at the previous time step k to the state at the current step k+1 in the absence of either a driving function or process noise. B_{k+1} is an n-by-m control input matrix that relates m-by-l control vector u_{k+1} to the state x_k. w_k is an n-by-1 system's noise vector with an n-by-n covariance matrix Q_k. $p(w_k) \sim N(0, Q_k)$. The measurement vector z_k contains state variables that are measured by the instruments. H_k is a j-by-n observation model matrix which maps the state x_k into the measurement vector z_k. v_k is a measurement noise j-by-1 vector with covariance R_k. $p(v_k) \sim N(0, R_k)$.

The Discrete Kalman filter has two distinct phases that compute the estimate of the next system's state [31].

Predict:

Predict the state vector ahead:

$$\hat{x}_{k+1}^{-} = A_{k+1} x_k + B_{k+1} u_{k+1} \tag{5}$$

The $\hat{x}_{k+1}^{-}$ is a prediction of the position of the future gaze.

Predict the error covariance matrix ahead:

$$P_{k+1}^{-} = A_{k+1} P_k A_{k+1}^T + Q_{k+1} \tag{6}$$

The predict phase uses the previous state estimate to predict the estimate of the next system's state.

Update:

Compute the Kalman gain:

$$K_{k+1} = P_{k+1}^{-} H_{K+1}^T (H_{k+1} P_{k+1}^{-} H_{k+1}^T + R_{k+1})^{-1} \tag{7}$$

Update the estimate of the state vector with a measurement z_{k+1}:

$$\hat{x}_{k+1} = \hat{x}_{k+1}^{-} + K_{k+1}(z_{k+1} - H_{k+1}\hat{x}_{k+1}^{-}) \quad \textbf{(8)}$$

Update the error covariance matrix:

$$P_{k+1} = (I - K_{k+1}H_{k+1})P_{k+1}^{-} \quad \textbf{(9)}$$

5.2 Two State Kalman Filter Model

The TSKF models an eye as a system with two states: position and velocity. The acceleration of the eye is modeled as white noise with known maximum acceleration and used in the design of the Q_k matrix for the Equation (6).

The TSKF models an eye as a system which has two state vectors x_k and y_k.

$$x_k = \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} \quad \textbf{(10)}$$

where $x_1(k)$ is horizontal coordinate of the gaze position and $x_2(k)$ is horizontal eye-velocity at time k.

$$y_k = \begin{bmatrix} y_1(k) \\ y_2(k) \end{bmatrix} \quad \textbf{(11)}$$

where $y_1(k)$ is vertical gaze position and $y_2(k)$ is vertical eye-velocity at time k.

The state transition matrix for both horizontal and vertical states is:

$$A = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \quad \textbf{(12)}$$

where Δt is the eye-tracker's eye-position sampling interval.

The observation model matrix for both state vectors is:

$$H = [1 \quad 0] \quad \textbf{(13)}$$

By definition the covariance matrix for the measurement noise is $R_k = E[(v_k - E(v_k))(v_k - E(v_k))^T]$. Because only eye position is measured v_k is a scalar making $R_k = VAR[v_k] = \delta_v^2$, where δ_v is the standard deviation of the measurement noise. In this paper, it is assumed that the standard deviation of the measurement noise relates to the accuracy of the eye tracker and is bounded by one degree of the visual angle. Therefore δ_v was conservatively set to 1°. In cases when the eye tracker fails to detect eye position coordinates, the standard deviation of measurement noise is assigned to be $\delta_v = 120°$. The value of 120° is chosen empirically, allowing the Kalman Filter to rely more on the predicted eye position coordinate $\hat{x}_k^-$.

The TSKF is initialized with zero valued initial vectors $\hat{x}_0$, $\hat{y}_0$ and an identity error covariance matrix P_0.

Parameterization of the process noise covariance matrix Q_k will result in three separate ROI construction models (*TSKF Simple*, *TSKF Content Accel*, *TSKF Fixed Accel*) with details presented in the next three subsections.

5.3 TSKF Simple

By definition, the process noise covariance matrix is $Q_k = E[(w_k - E(w_k))(w_k - E(w_k))^T]$, where w_k is a 1x2 system's noise vector $w_k = [w_1(k) \quad w_2(k)]^T$. *TSKF Simple* model assumes that variables $w_i(k)$ are uncorrelated between each other (velocity is independent of eye position), i.e., $E[(w_m(k)w_n(k)] = E[(w_m(k)]E[w_n(k)]$ for all $n \neq m$ and $p(w_1(k)) \sim N(0, \delta_1^2)$, $p(w_2(k)) \sim N(0, \delta_2^2)$. These assumptions generate following the system's noise covariance matrix: $Q_k = \begin{bmatrix} \delta_1^2 & 0 \\ 0 & \delta_2^2 \end{bmatrix}$. The *TSKF Simple* model assumes that the standard deviation of the eye position noise $w_1(k)$ is connected to the characteristics of the eye fixation movement. Each eye fixation consists of three basic eye-sub-movements: drift, small involuntary saccades, and tremor [32]. Among those three, involuntary saccades have the highest amplitude -

about a half degree of the visual angle; therefore δ_1 is set conservatively to 1°. Standard deviation value for eye velocity was selected to be $\delta_2 \ = 1$°/s.

5.4 TSKF Acceleration-Based Approaches

Kohler has derived a process noise covariance matrix Q_k for a system that models translational motion with constant velocity and white noise acceleration of the maximum amplitude a [33]. Suggested Q_k matrix was:

$$Q_k = \frac{a^2 \Delta t}{6} \begin{pmatrix} 2I(\Delta t)^2 & 2I(\Delta t) \\ 3I(\Delta t) & 6I \end{pmatrix} s \qquad \textbf{(14)}$$

where I is an identity matrix; $\Delta t = t_{k+1} + t_k$ – time between eye-position sampling, s - unit seconds; a – acceleration.

Out of the three eye movements, fixation, pursuit, and saccade only, saccades are responsible for the rapid eye rotation and exhibit the highest acceleration values. Therefore, it would be reasonable to assume that properties of the saccadic eye movements should define acceleration parameter a for the matrix Q_k. Two approaches define the value for a. The first approach, the *TSKF Content Accel*, monitors eye acceleration during stimuli perception and assigns the highest recorded eye acceleration value to a. The second approach, *TSFK Fixed Accel*, takes into account that that high average eye acceleration is approximately 4000°/s^2 [10] and assigns this fixed value to a.

5.5 ROI Construction

For each ROI placement model, i.e., *TSKF Simple*, *TSKF Content Accel*, *TSKF Fixed Accel* two separate instances of the given model are employed in the eye movement prediction process. The first instance is the base which provides updated gaze coordinates each time a measurement arrives from an eye tracker. The second instance is the actual predictor that uses updated coordinates as an input and predicts future gaze position given the feedback delay value. The ROI center is placed on top of the predicted gaze. A previous history of the magnitude of the error between predicted and measured eye position determines the size of the ROI. The following paragraphs provide description that is more detailed.

The TSKF considers the horizontal and the vertical components of the eye movement separately. The base model iterates each time a new measurement arrives from the eye tracker providing updated gaze coordinates and covariance error matrixes ($\hat{x}_{k+1}$, P_{k+1} and $\hat{y}_{k+1}$, P_{k+1}). Those values are supplied as initial values to the new instance of the TSKF that predicts the coordinates of the future gaze, providing the compensation for the delay. The number of iterations required for prediction can be represented by the value $\mathrm{m} = \lceil \mathrm{T_d} \cdot \mathrm{eye\_tracker\_sampling\_frequency\_HZ} \rceil$. The challenge presented in front of the predictor TSKF is to obtain the position of the future gaze samples given the value of the feedback delay with no measurements present during this time interval. The challenge is resolved by assigning the required measurements $\mathrm{z_k}$ for horizontal and vertical components to $\hat{x}_{k+1}$and $\hat{y}_{k+1}$ mentioned above. Very importantly, the standard deviation of the measurement noise δ_v is assigned to be 120° making the measurement noise covariance matrix $R_k = 120^2$. This assignment allows the Kaman filter to rely more on the HVS model for prediction and practically ignore the last valid eye position measurement in the update stage. Theoretically, the δ_v should be set to infinity; but practically, the value of $\delta_v = 120$ brings better results. As a result of m iterations of the predictor TSKF, the values $\hat{x}_{k+m+1}^{-}$ and $\hat{y}_{k+m+1}^{-}$, become the predicted coordinates of the future gaze. The ROI center coordinates become ($x_{cen}^{ROI}(k) = \hat{x}_{k+m+1}^{-}$, $y_{cen}^{ROI}(k) = \hat{y}_{k+m+1}^{-}$).

The ROI dimensions are selected to be proportional to the gaze position prediction error. Ideally, such an error would be normally distributed $N(\mu, \sigma^2)$ with $\mu = 0$. The relationship between the amount of samples contained under the bell curve of the normal distribution between $(\mu - n\sigma^2, \mu + n\sigma^2)$ can be represented by the integral [34]

$$F(n) = \frac{2}{\sqrt{\pi}} \int_{0}^{n/\sqrt{2}} e^{-t^2} dt \qquad \textbf{(15)}$$

Therefore, if the ROI dimensions are assigned to be $(n\sigma_x^2, n\sigma_y^2)$ the solution to the equation TGC=F(n) will provide the value for n, necessary to contain the amount of gaze samples specified by TGC. Unfortunately, integral (15) cannot be expressed in terms of elementary functions; therefore, linear approximation n = (TGC-41)/27 is employed. This linear approximation is based on the fact that approximately 68% of values drawn from a normal distribution are within one standard deviation from the mean, and 95% of the values are within two standard deviations [34]. The use of the linear approximation of n value will not be advisable for the TGC greater than 95% due to inaccuracies appearing as a result of the approximation.

Standard deviation (SD) of the prediction error is calculated as $\sigma_x(t) = \sqrt{\sum_{k=j}^{t-T_d} \frac{\left(\hat{x}_{k+m+1}^{-} - x_{measured}\right)^2}{(t-T_d)\cdot\frac{AEGC\ (t)}{100}}}$, where AEGC(t) is the average amount of gaze samples contained by the ROI from the moment j of the recording up to the moment t. The value t-T$_d$-j will represent the size of the sampling window where the calculation of $\sigma_x(t)$ and the AEGC(t) occurs. The existence of such a temporary window will prevent having "outdated" samples for $\sigma_x(t)$. For more accurate results, the size of this window should be content dependent. In the experiments presented in this paper, j was set to 1. The maximum possible value for AEGC(t) is 100. The performance of the real-time *Gaze-contingent Compression System* (GCS) has to be conservative, i.e., actual gaze containment has to be above the targeted gaze containment to provide the required perceptual quality. The value applied in denominator $\frac{AEGC(t)}{100}$ increases $\sigma_x(t)$ in cases when the actual gaze containment goes down, therefore increasing the size of the ROI and containing more gaze samples. The same logic is applied in calculating the standard deviation for the vertical component of movement.

It can be argued that the suitable value for $\sigma_x(t)$ can be extracted from the error covariance matrix P_k calculated by the Kalman filter, but the investigation of the approach is beyond the scope of this paper.

6 EXPERIMENT SETUP

6.1 Eye Movement Detection Algorithm

It is important to estimate the performance of the ROI construction in terms of various eye movement types such as fixations, saccades, and pursuits. This work adopts the *Velocity-Threshold Identification* (I-VT) model [35]. The original I-VT model proposed by Salvucci & Goldberg classified eye position samples as fixation when eye velocity was below 100°/s and saccades when eye velocity was above 300°/s. These values seem to be in contradiction with oculomotor research and neurological literature. Leigh and Zee [30] indicate that saccade onset is detected when eye velocity rises to 30°/s, and smooth pursuit eye movements are detected with eye velocities of 5-30°/s. A different group of researchers represented by Meyer and colleagues *[36]* found out that the *human visual system* (HVS) maintains pursuit motion with velocities of up to 90°/s. These facts show that it is hard to define an automated instant eye movement detection criterion, especially if eye position sample data is the only source of information. Nevertheless, this paper adopts classification based on velocity values suggested by Leigh and Zee: fixations (0-5°/s), saccades (more than 30°/s), and pursuits (5-30°/s).

6.2 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 media 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. Several video clips were looked at, each offering different combinations of subjective complexities. In this paper, three representative cases are considered with rough subjective-complexity descriptions presented below:

Car: This video shows a moving car. It was taken from a security camera view point 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 movements. Several pedestrians and distant cars appeared in the background several times, often capturing the attention of the subject. The remaining part of the video's background was still.

Shamu: This video captures a night performance of an Orca whale under a tracking spotlight. The video consists of several moving objects: the whale, the trainer, and the crowd. Each of those objects 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 results in a variety of eye movements: fixations, saccades, and smooth pursuits. The background of the video was constantly in motion due to the fact that the camera was trying to follow the swimming whale. Such a stimulus suits the goal of challenging prediction models 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 by a subject.

Airplanes: This video depicts formation flying of supersonic planes, rapidly changing their flying speeds. The number of planes varies from one to five during the clip. The scene-recording camera movements were rapid zoom and panning. Frequently the camera could not focus very well on a plane and the subject had to search for it. This aspect brought an additional complication to the general pattern of eye movements. The background of this video was in constant motion and presented a blue sky.

All three videos had a resolution of 720x480 pixels, frame-rate of 30fps, and were between 1 and 2 minutes long [37].

6.3 Equipment & Setup

The experiments were conducted with a Tobii 1750 eye tracker which is represented by a 17 inch flat panel with resolution of 1280x1024 and built-in eye tracking camera. This eye tracker performs binocular tracking with the following characteristics: sampling rate - 50Hz, accuracy 0.5°, spatial resolution 0.25°, and drift less than 1°. The Tobii 1750 model allows 300x150x200 mm freedom for the head movement [38]. Nevertheless, during the experiments, every subject was asked to hold his/her head motionless. Subjects were seated approximately 650 mm from the eye tracker. The experiment facilitator monitored subjects during each recording for any head movements outside of 300x150x200 mm range. No such movements were reported. Before running each experiment, the eye tracker was calibrated for each subject and checked for calibration accuracy. The stimuli were presented in the following order for each subject "Car", "Shamu", and "Airplanes". None of the video files was available publicly; therefore, it was safe to assume that this was the first time the subjects saw these stimuli.

The performance evaluation was conducted for a 0.02-2 s delay range and the TGC of 60-90%. All performance analysis was done off-line with feedback delay simulated as a time difference between predicted and measured gaze samples T_d seconds apart.

6.4 Participants

The subject pool consisted of 21 volunteers of both genders and mixed ethnicities, aged 20-40 with normal, corrected and uncorrected vision. Subjects with uncorrected vision reported that they were comfortable working with a computer without vision correction. The subjects were instructed to watch the video clips in any way they wanted.

6.5 Evaluation Metrics

6.5.1 Gaze Containment

Average Gaze Containment (AEGC) metric computes the number of gaze samples that are contained within the ROI region. The AEGC metric can be considered as a quantitative measurement of the perceived quality of the gaze-contingent compression system – the more gaze samples are contained within the high quality coded ROI, the lesser is the probability that compression artifacts are detected. Mathematically AEGC is computed as:

$$AEGC = \frac{100}{N}\sum_{k=1}^{N} GAZE^{ROI}(k) \qquad \textbf{(16)}$$

Variable $GAZE^{ROI}(k)$ equals one when k^th gaze falls within the ROI boundaries and zero otherwise. N is the number of gaze samples under consideration. AEGC metric can be employed to measure gaze containment during fixations, saccades, and pursuits. For the perceptually lossless gaze-contingent compression systems, AEGC should be equal to 100% with the ideal ROI size equivalent to the size of the fovea. In practice, AEGC should be equal to the TGC parameter set by the system designer.

6.5.2 Perceptual Resolution Gain

The actual amount of bandwidth/computational burden reduction achieved by a gaze-contingent compression system will depend on the ROI size and the method selected for peripheral image degradation [7]. The compression estimation approach selected in this paper is based on the visual sensitivity function presented in Section 4.3 and is calculated by a metric *Average Perceptual Resolution Gain* (APRG) computed with the formula:

$$APRG = \frac{M \cdot H \cdot W}{\sum_{k=1}^{M} \int_0^W \int_0^H S_k(x,y)dxdy} \qquad \textbf{(17)}$$

where M represents a set of gaze samples during an experiment. $S_k(x,y)$ is the eye sensitivity function defined by Equation (2). The ROI size required for the $S_k(x,y)$ calculation is extended by one degree in every direction to reduce the probability of compression detection in cases when gaze falls on the ROI boundary. W and H are the width and height of the visual image.

The APRG values above one will indicate that the gaze-contingent approach provides additional savings in terms of bandwidth/computational burden reduction. The APRG value of one will indicate that the gaze-contingent approach with given parameters does not provide any benefits. It is important to understand the maximum benefit that gaze contingent compression can achieve. It is assumed that such case occurs when feedback delay is zero and the eye-gaze is directed toward the center of the screen for the duration of the experiment. According to the Equation (2) and assuming that the ROI's radius is 1° of the visual angle and taking into consideration setup parameters presented in Section 6 the maximum value for the APRG can be estimated to be 2.6.

7 RESULTS

The experiment results which show the performance of the gaze-contingent compression system with incorporated delay compensation algorithm are presented by Figure 4 and Figure 5. The results were averaged between all recordings.

Between three models, the *TSKF Simple* provided the highest compression results (APRG of 2.5-1.4) with the AEGC values higher than the TGC values for small feedback (0.02 s). For larger feedback delays (0.5-2 s), the AEGC was lower than the TGC but not more than by 7%.

The *TSKF Fixed Accel* was the most conservative model in terms of the AEGC performance given the value of the TGC, i.e., the AEGC value was always higher than the TGC value. This property was achieved at the cost of reduced compression (0-18% reduction) for the same input parameters.

The performance of the *TSKF Content Accel* was similar to the *TSKF Simple* model for small and medium delays (0.02-0.5 s), but in the case of large delays (1-2 s) the *TSKF Content Accel* was able to provide a 4% higher AEGC while achieving a 5-6% higher compression, indicating the benefit of eye gaze-based acceleration choice. It should be noted that for large delays, the AEGC values achieved by the *TSKF Content Accel* model were lower than the targeted values by up to 3%; therefore, the *TSKF Fixed Accel* model was still the most conservative choice with more gaze containment but slightly less compression.

8 DISCUSSION

8.1 Perceptual Quality and AEGC

This paper did not evaluate subjectively the distraction impact caused by the gaze samples not being contained by the ROI, i.e., gaze containment calculation and compression estimation were done off-line. Therefore, it is important to ask how "bad" it would be for a real-time gaze-contingent compression system if the gaze sample belonging to an eye fixation or pursuit falls outside of the high quality coded ROI. The results might vary, depending on how far the missed eye-fixation is from the ROI boundary and the implementation of the visual sensitivity function. In case a viewer notices the "blurred" effect and is unable to see a specific detail in the picture, he or she can fixate the eyes on the point in question; and the system will place the ROI on the spot under attention. The amount of time required for this operation will directly depend on the delay duration. Additionally, it is important to take into consideration that the timing of the ROI placement does not have to be perfect, i.e., if the ROI is placed 60ms after the onset of a fixation, the system remains perceptually lossless [14]. Further investigation is necessary to match gaze containment numbers to specific goals imposed on a gaze-contingent compression system.

8.2 TSKF vs HESA

The *Histogram Eye Speed Analysis* (HESA) model was introduced by Komogortsev and Khan in [15]. The HESA model is a target gaze containment performance oriented model. The HESA model places the ROI center on the last available to the system gaze sample with the ROI dimensions defined as a product of the *Future Predicted Eye Speed* (FPES) and the feedback delay value. FPES is computed by a histogram analysis of the recorded eye speeds.

8.2.1 Gaze Containment

One of the weaknesses of the HESA model was the fact that the model required empirical selection of the TGC parameter to receive the required AEGC. For the TGC=90%, the AEGC values were close to 100%, and no compression was achieved. Empirically selected TGC values allowing the AEGC achievement of 90% were 18-44% below the 90% mark. As it was pointed out in Section 7, the most conservative model, the *TSKF Fixed Accel*, in instances where TGC=90%, the provided AEGC values were always higher than the TGCwith the maximum recorded difference of 2-6% (0.02s delay case). Therefore, it is possible to conclude that the *TSKF Fixed Accel* model provides a more robust performance in terms of the AEGC, given the value of the TGC, than the HESA model.

8.2.2 Compression

To compare the performance of the TSKF based models and the HESA model, the *TSKF Fixed Accel* model was selected as the model providing the most conservative performance in terms of AEGC vs. TGC. The following table presents APRG values obtained by the HESA and the *TSKF Fixed Accel* for 0.5 – 2 s delay and the AEGC of 90% or higher.

Model Name/Delay value	*0.5 s*	*1 s*	*2 s*
HESA	1.19	1.14	1.12
TSKF Fixed Accel	1.5	1.44	1.4
One-Way ANOVA	F(1, 64)=3.46, p=0.067	F(1,64)=3.7, p=0.059	F(1,64)=3.9, p=0.052

It is possible to conclude that, on average, the *TSKF Fixed Accel* provided more compression with a strong trend toward statistical significance. Tests with a higher number of subjects have to be performed to establish results that are more accurate.

8.3 Average Gaze Containment vs. Average Fixation Containment

This paper continues the empirical evaluation of the hypothesis that the Average Gaze Containment is a more conservative metric than the Average Eye Fixation Containment (AFC). Originally this hypothesis was proposed by Komogortsev and Khan [15] and was empirically validated for the Dwell-Time Fixation detection method in the same work.

Experiments conducted with the I-VT eye movement detection model for this paper indicate that the hypothesis is true for low feedback delay cases. More specifically, in the case of the "Car" and "Airplanes" videos, the AFC was always higher than the AEGC. For small feedback values of 0.02 s, the difference between the AFC and the AEGC was statistically significant (pair-wise t-test, $p<0.05$) for all models and TGC values and not statistically significant for feedback delay values of 0.5 s and higher. In the case of the "Shamu" video, the AFC was always higher than the AFC for the delays of 0.02 s ($p<0.05$). For higher delay cases, the AFC was lower in several instances than the AEGC, but the difference did not exceed 4% (not statistically significant).

The *Average Pursuit Containment* (APS) was higher than the AEFC in all scenarios. The *Average Saccade Containment* (ASC) was less than the AEGC in all cases. The HVS is effectively blind during saccades [10]; therefore, low ASC should not negatively affect the performance of a gaze-contingent compression system.

Figure 6 depicts two representative cases of the average gaze containment for all movement types for the "Car" and "Shamu" videos.

9 CONCLUSION

This paper presented a Kalman filter-based algorithm for the sensor/processing/transmission delay compensation in a gaze-contingent compression system with the targeted gaze containment performance. The algorithm constructed an elliptically shaped high quality coded ROI and placed this ROI on top of the predicted gaze coordinates. The goal of the ROI was to contain a targeted amount of gaze samples within the smallest ellipse, given the value of the delay. The ROI parameters were calculated by a two state Kalman filter that modeled an eye as a system with translational motion and white noise acceleration. Content-based and fixed acceleration selection approaches were employed. The ROI construction was evaluated by 21 subjects, three multimedia files, 0.02-2 s delay range, and such metrics as: gaze containment (quantitative estimation of the quality of the compressed image) and perceptual resolution gain (estimation of bandwidth/computational burden reduction achieved by gaze-contingent compression). The performance of the gaze contingent compression system was assessed with the targeted gaze containment of 60-90%. Content-based and low acceleration parameterization provided the compression of 1.4-2.5 but for some of the high delay cases, average gaze containment was 7% less than the targeted value. The model with fixed eye acceleration of 4000°/sec^2 always provided average gaze containment that was higher than targeted values with a reported compression of 1.4-2.3 times. It is possible to hypothesize that a more accurate model of the human visual system [39] should improve the predictive performance of the Kalman filter and, therefore, bring higher compression results.

Gaze containment and perceptual resolution gain numbers achieved by the TSKF model can be compared to the previously published *Histogram Eye-Speed Analysis* (HESA) model [15] that achieved 90% of the average gaze containment and 1.12-1.19 compression for the 0.5-2 s delay range. The most conservative Kalman filter-based model presented in this work achieved approximately a 25% higher compression with the same perceptual performance.

The experiments conducted in this paper partially support the hypothesis suggested by Komogortsev and Khan [37] that gaze containment is a more conservative metric for the evaluation of the perceptual quality of the gaze-contingent compression than a fixation-containment metric. The results indicate that the average gaze containment metric can be considered as a lower boundary for average fixation/pursuit containment and an upper boundary for the average saccade containment.

Additional research should be performed to evaluate the effect of distraction caused by the gaze samples that are not contained by the ROI. Different levels of gaze containment have to be mapped to a set of goals imposed on a gaze-contingent compression system.

10 GLOSSARY

AEGC – Average Eye Gaze Containment

AFC – Average Fixation Containment

APC- Average Pursuit Containment

ASC – Average Saccade Containment

GCS - Gaze-contingent video Compression System

HVS - Human Visual System

ROI – region of interest

TGC – Targeted eye Gaze Containment

TSKF – Two State Kalman Filter

11 BIBLIOGRAPHY

1. D. E. Irwin, "Visual Memory Within and Across Fixations," Eye movements and Visual Cognition: Scene Preparation And Reading 146-165 (1992)

2. P. Kortum and W. Geisler, S., "Implementation of a foveated image coding system for image bandwidth reduction," in Proceedings of SPIE: Human Vision and Electronic Imaging, pp. 350-360 (1996).

3. W. Geisler, S. and J. Perry, S., "Real-time foveated multiresolution system for low-bandwidth video communication," in Proceedings of SPIE: Human Vision and Electronic Imaging III, pp. 294-305 (1998).

4. A. Duchowski, T. and B. McCormick, H., "Gaze-contingent video resolution degradation," in Proceedings of SPIE: Human Vision and Electronic Imaging III, pp. 318-329 (1998).

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

6. S. Lee, M. Pattichis and A. Bovok, "Foveated Video Compression with Optimal Rate Control," IEEE Transactions of Image Processing 10(7), 977-992 (2001)

7. O. Komogortsev, V. and J. Khan, I., "Predictive Perceptual Compression for Real Time Video Communication," in 12th ACM International conference on Multimedia, pp. 220-227, New York (2004).

8. O. Komogortsev, V. and J. Khan, "Perceptual Multimedia Compression based on the Predictive Kalman Filter Eye Movement Modeling," in Multimedia Computing and Networking Conference (MMCN'07), pp. 1-12, San Jose (2007a).

9. H. Murphy and A. Duchowski, T., "Hybrid Image-/Model-Based Gaze-Contingent Rendering," in Applied Perception in Graphics and Visualization, ACM, Tuebingen (2007).

10. A. Duchowski, Eye Tracking Methodology: Theory and Practice Springer (2007).

11. L. Loschky, C. and G. McConkie, W., "User performance with gaze contingent multiresolutional displays," in Symposium on Eye Tracking Research & Applications, pp. 97-103, New York (2000).

12. L. Sanghoon, M. Pattichis, S. and A. Bovik, C., "Foveated video compression with optimal rate control," IEEE Transactions on Image Processing 10(7), 977-992 (2001)

13. W. Zhou and B. A. C., "Embedded Foveation Image Coding," IEEE Transactions on Image Processing 10(10), 1397-1410 (2001)

14. L. Loschky, C. and G. Wolverton, S., "How late can you update gaze-contingent multiresolutional displays without detection?," ACM Transactions on Multimedia Computing, Communications, and Applications 3(4), (2007)

15. O. V. Komogortsev and J. I. Khan, "Predictive real-time perceptual compression based on eye-gaze-position analysis," ACM Trans. Multimedia Comput. Commun. Appl. 4(3), 1-16 (2008)

16. D. Parkhurst, J. and E. Niebur, "Variable resolution displays: A theoretical, practical, and behavioral evaluation," Human Factors 44(4), 611-629 (2002)

17. E. Reingold, M., L. Loschky, C., G. McConkie, W. and D. Stampe, M., "Gaze-Contingent multi-resolutional displays: An integrative review," Human Factors 45(2), 307-328 (2003)

18. A. Duchowski, T., N. Cournia and H. Murphy, "Gaze-Contingent displays: A review.," Cyberpsychological Behavior 7(6), (2004)

19. A. Duchowski, T. and A. Çöltekin, "Foveated gaze-contingent displays for peripheral LOD management, 3D visualization, and stereo imaging.," ACM Transactions on Multimedia Computing, Communications, and Applications 3(4), (2007)

20. A. Duchowski, T., "Acuity-matching resolution degradation through wavelet coefficient scaling," IEEE Transactions on Image Processing 9(8), 1437-1440 (2000)

21. T. Kuyel, W. Geisler and J. Ghosh, "Retinally reconstructed images (RRIs): digital images having a resolution match with the human eye," in Proceedings of SPIE: Human Vision and Electronic Imaging III, pp. 603-614 (1998).

22. L. Itti and C. Koch, "Computational Modelling of Visual Attention," Neuroscience 2(1-11 (2001)

23. R. Carmy and L. Itti, "Casual Saliency Effects During Natural Vision," in Symposium on Eye Tracking Research & Applications 2006 (ETRA 06), pp. 11-18, San Diego (2006).

24. R. Peters and L. Itti, "Computational mechanism for gaze direction in interactive visual environments," in Symposium on Eye tracking Research & Applications 2006 (ETRA 06), pp. 27-32, San Diego (2006).

25. A. Duchowski, T. and B. McCormick, H., "Preattentive considerations for gaze-contingent image processing," in Proceedings of SPIE: Human Vision, Visual Processing, and Digital Display VI, pp. 128-139 (1995a).

26. A. Duchowski, T. and B. McCormick, H., "Simple multiresolution approach for representing multiple regions of interest (ROIs)," in Proceedings of SPIE: Visual Communications and Image Processing, pp. 175-186 (1995b).

27. J. Khan and O. V. Komogortsev, "A Hybrid Scheme for Perceptual Object Window Design with Joint Scene Analysis and Eye-Gaze Tracking for Media Encoding based on Perceptual Attention," Journal of Electronic Imaging 15(2)(023018-023001 - 023018-023002 (2006)

28. O. V. Komogortsev and J. Khan, "Perceptual Attention Focus Prediction for Multiple Viewers in Case of Multimedia Perceptual Compression with Feedback Delay," in Symposium on Eye Tracking Research & Applications 2006 (ETRA 06), pp. 101-108, San Diego (2006).

29. O. V. Komogortsev and J. Khan, "Kalman Filtering in the Design of Eye-Gaze-Guided Computer Interfaces," in 12th International Conference on Human-Computer Interaction (HCI 2007), pp. 1-10, Beijing, China (2007b).

30. R. J. Leigh and D. S. Zee, The Neurology of Eye Movements, Oxford University Press (2006).

31. R. Brown and P. Hwang, Introduction to Random Signals and Applied Kalman Filtering, John Wiley and Sons, New York (1997).

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

33. M. Kohler, "Using the Kalman Filter to track human interactive motion --- Modelling and initialization of the Kalman Filter for translational motion," (1997).

34. W. Feller, Ed., An Introduction to Probability Theory and Its Applications, Wiley (1968).

35. D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye tracking protocols," in Eye Tracking Research and Applications Symposium, pp. 71-78, ACM Press, New York (2000).

36. C. H. Meyer, A. G. Lasker and D. A. Robinson, "The Upper Limit of Human Smooth Pursuit Velocity," Vision Research 25(4), 561-563 (1985)

37. O. V. Komogortsev and J. I. Khan, "Eye movement prediction by Kalman filter with integrated linear horizontal oculomotor plant mechanical model," in Proceedings of the 2008 symposium on Eye tracking research and applications, ACM, Savannah, Georgia (2008).

38. Tobii, "Tobii 1750 Eye-tracker," Tobii Technology (2003).

39. O. V. Komogortsev and J. Khan, "Eye Movement Prediction by Oculomotor Plant Kalman Filter with Brainstem Control," Journal of Control Theory and Applications 7(1), (2009)

12 LIST OF FIGURE CAPTIONS

Error! Reference source not found.

Error! Reference source not found.

Error! Reference source not found.

Error! Reference source not found.

Error! Reference source not found.

Error! Reference source not found.

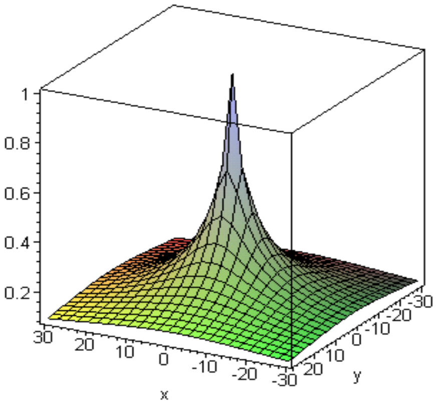

Figure 1. Visual Sensitivity Function: single gaze point case.

(

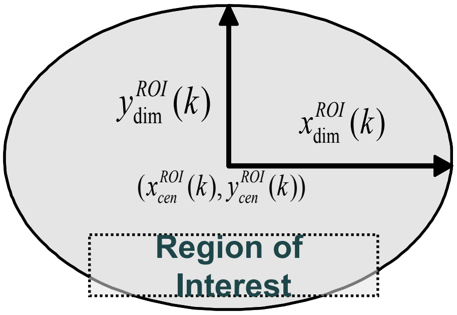

Figure 2. Region of Interest: dimensions are designed to compensate for the delay effects.

(

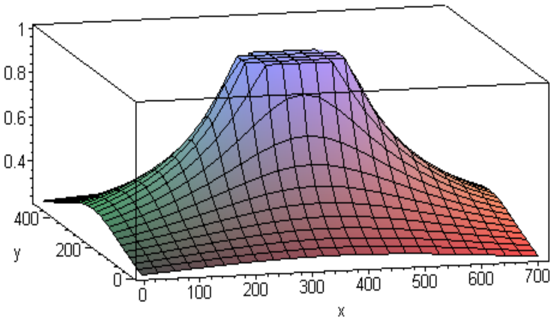

Figure 3. Visual Sensitivity Function: with incorporated ROI.

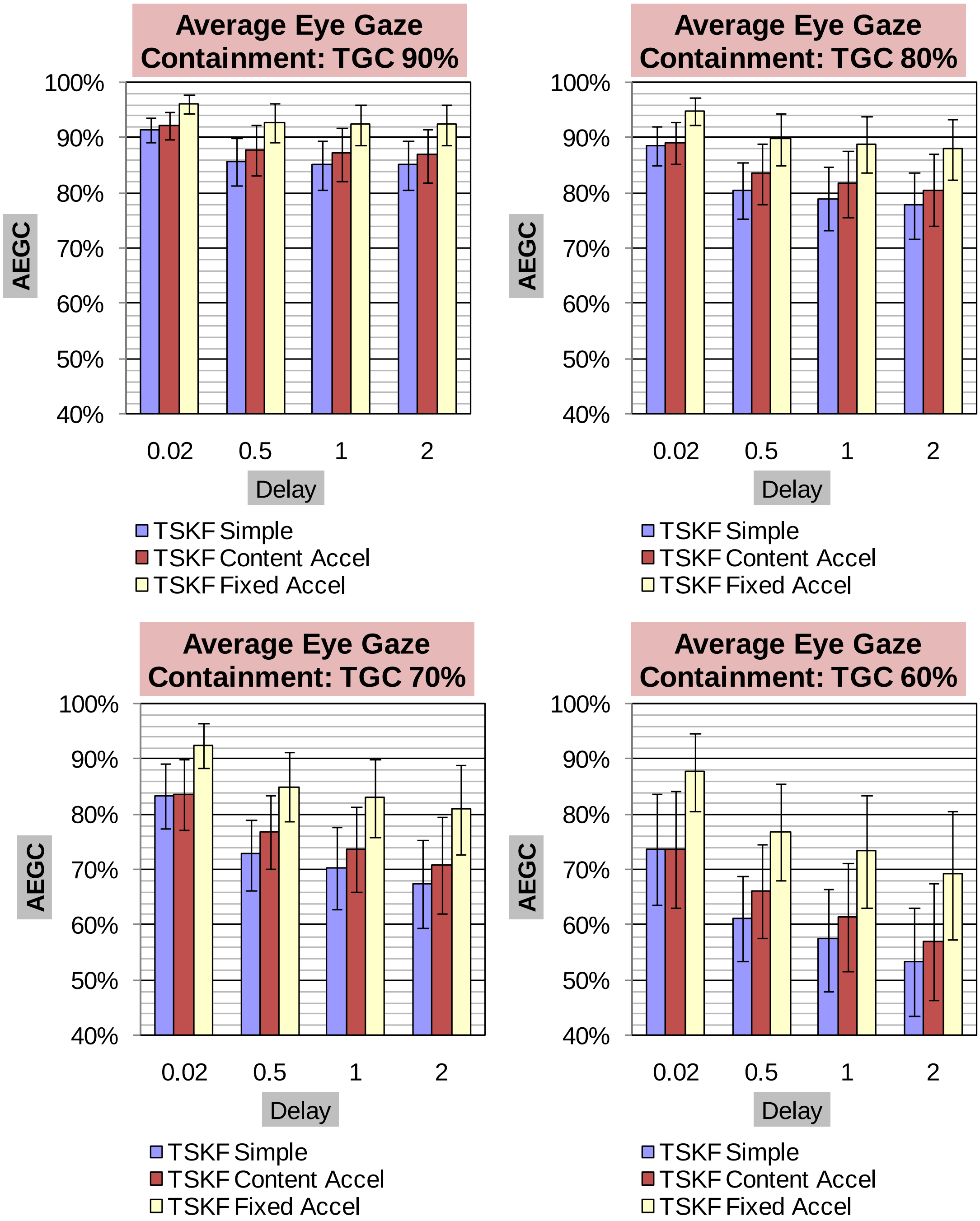

Figure 4. Average Eye Gaze Containment performance for various target gaze containment and delay scenarios. Error bars represent standard deviation.

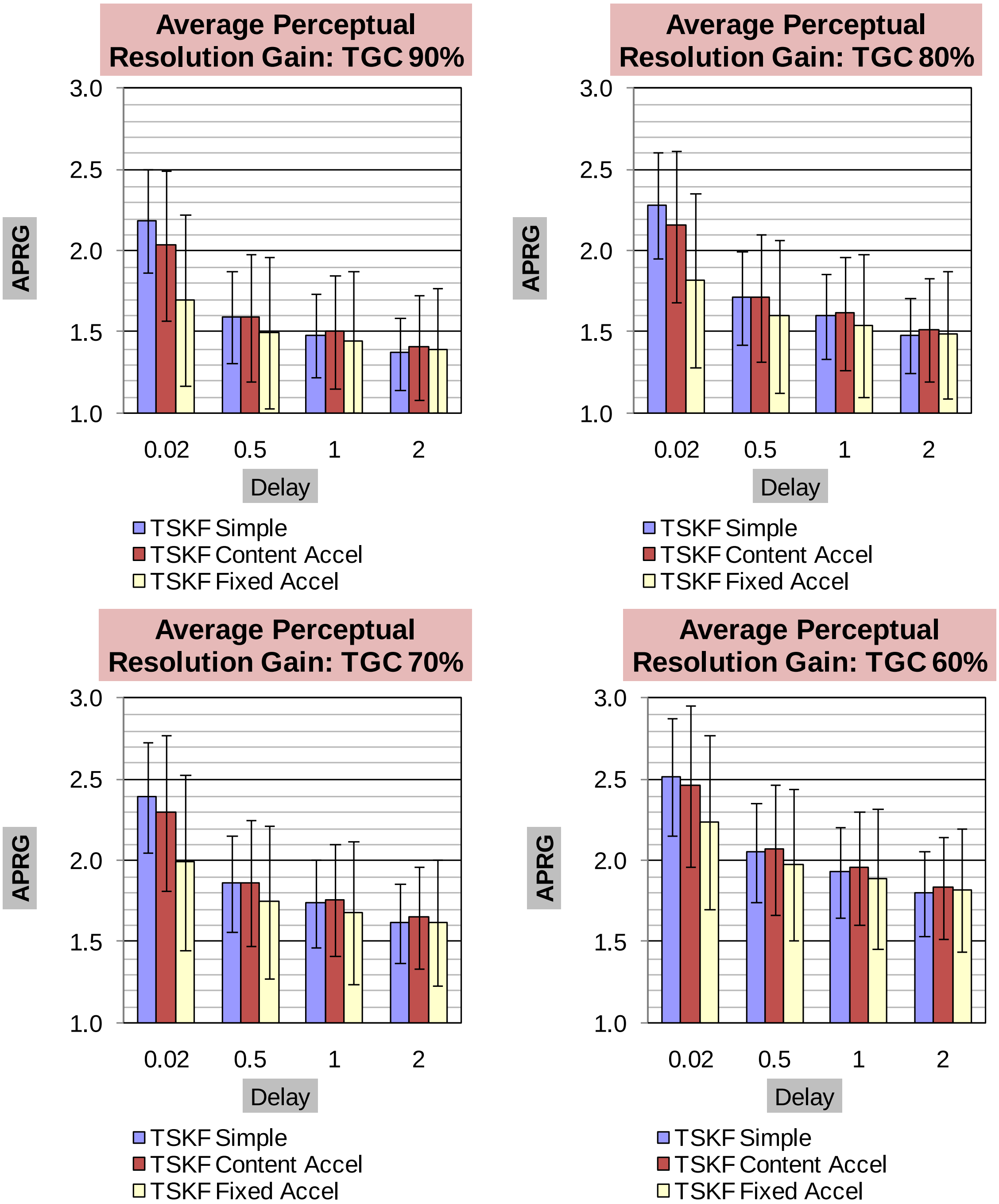

Figure 5. Average Perceptual Resolution Gain performance for various target gaze containment and delay scenarios. Error bars represent standard deviation.

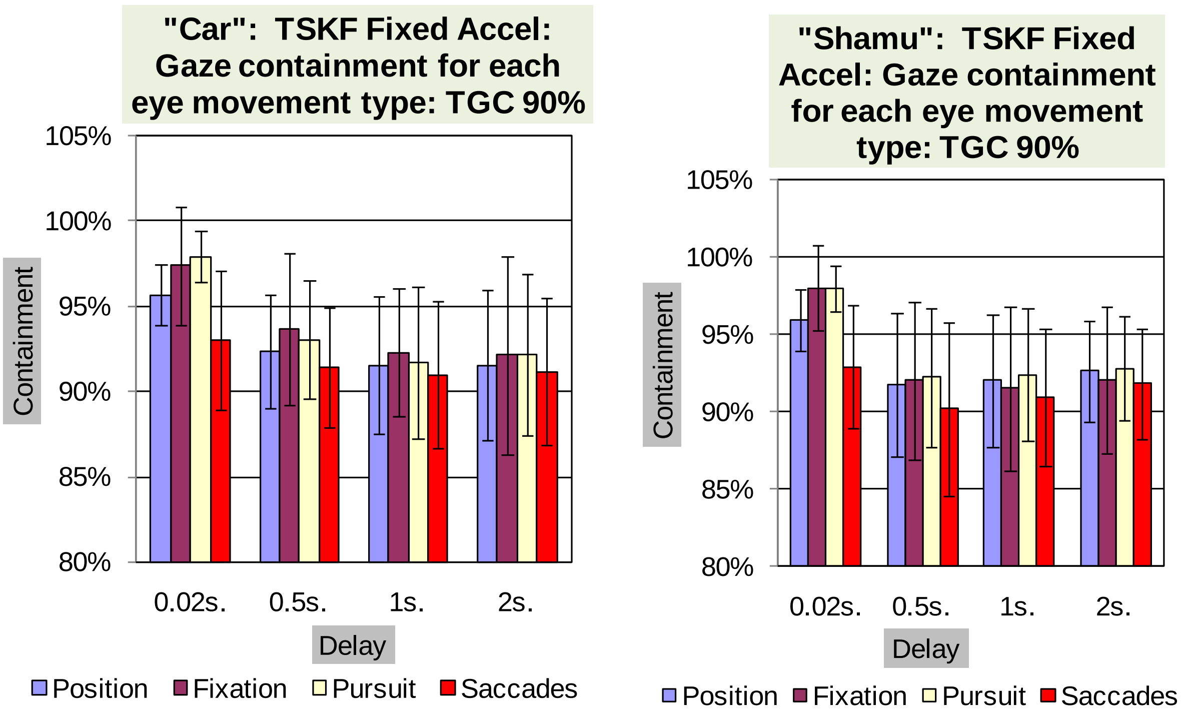

Figure 6. Average Gaze vs. Fixation vs. Pursuit vs. Saccade Containment.