A Hybrid Scheme for Perceptual Object Window Design with Joint Scene Analysis and Eye-Gaze Tracking for Media Encoding based on Perceptual Attention

Javed I. Khan and Oleg Komogortsev
Media Communications and Networking Research Laboratory
Department of Math & Computer Science, Kent State University
233 MSB, Kent, OH 44242

ABSTRACT

The possibility of perceptual compression using live eye-tracking has been anticipated for some time by many researchers. Among the challenges of real-time eye-gaze based perceptual video compression is how to handle the fast nature of eye movements with a relative complexity of video transcoding and also take into the account a delay associated with transmission in the network. Such delay requires an additional consideration in perceptual encoding because it increases the size of the area that requires high quality coding. In this paper we present a hybrid scheme, one of the first to our knowledge, which combines eye-tracking with fast in-line scene analysis to drastically narrow down the high acuity area without the loss of eye-gaze containment.

Keywords: eye-gaze, perceptual encoding, MPEG-2.

1. INTRODUCTION

Perceptual coding is emerging as a new area of high fidelity media coding. It is believed that our eye sees only a part of the visual plane at any given time. The intention of perceptual video compression is to calculate the spatial distribution of bits with close coherence of the perceptually meaningful shapes, objects and actions presented in a scene. A number of researchers have confirmed the effectiveness of this approach. However, there are many challenges in the engineering of such scheme. It is difficult to detect the perceptual significance of a particular area in a given video scene. Indeed, the nature of perception dictates that our previous experience plays an important role in the visualization process. Thus, a scene itself may not have all the information that guides visual attention. Researchers in vision and eye-tracking have suggested perceptual coding with direct eye-tracker based detection of visual attention. Only about 2 degrees in our about 140 degree vision span has sharp vision. A fascinating body of research exists in vision and psychology geared towards the understanding of the human visual system. These techniques process information from eye and head tracking devices and attempt to obtain the correct eye-gaze position with respect to a visual plane. Based on the acuity distribution characteristics of a human eye around the fovea these methods use variable spatial resolution coding for image compression. These techniques require precise spatio-temporal information about the position of the eye. In the case of a large network delay, or encoding delay, an eye can move away from its detected location by the time the information is received and processed. This severely offsets the 2 degree acuity advantage.

Recently, we have performed several experiments to develop a hybrid technique that combines direct eye gaze sampling with a video scene content analysis. It is believed that both scene content and an eye movement pattern determine the precise area of human attention. Our hybrid scheme uses both these facts to calculate the exact area of the image that requires high quality coding. The hybrid scheme first creates a *Reflex Window* (W^{RW}) based on eye-gaze information, determining where subject's visual attention is directed. Then it calculates an Object Window (W^{TW}) based on a fast content analysis algorithm, predicting the vicinity of subject's focus. After these two steps a new Perceptual Object Window (W^{POW}) is constructed based on the areas provided by W^{RW} and W^{TW} increasing the advantages and reducing disadvantages of each.

We show that our technique reduces the area requiring high quality coding, thus increasing the scope of compression. Also, our method enables more eye gazes to be contained within the W^{POW} for its size, thus retaining the perceptual quality. This technique is probably one of the first that merges the two major paradigms of perceptual encoding. The overall Perceptual Object Window is media independent and can be applied to any video compression method. We have also recently completed an MPEG-2 implementation of proposed hybrid scheme.

The two base techniques for scene analysis and eye-gaze based Reflex Window were described separately in [5] and [8]. In this paper we present them briefly in the following two sections. In section 4 we present several possible schemes for combining them. Then in section 5 we present experiment results and analysis.

1.1. Related Work

A large number of studies have been performed that investigated various aspects of perceptual compression. A particular focus has been the study of contrast sensitivity or spatial degradation models around the foveation center and its impact on the perceived loss of quality by subjects [2, 9, 11, 15]. [4] presented pyramid coding and used pointing device to identify focus. [6, 7] demonstrated mouse driven high resolution window overlay interface for medical video visualization over bandwidth constrained links. Many of the early works have been inspired by the objective to design a good quality display systems [1, 3, 11]. For example, [1] utilized a live eye tracker to determine the maximum frequency and spatial sensitivity for HDTV displays with fixed observer distance. [10] discussed how to optimally control the bit-rate for MPEG-4/ H.263 stream for foveated encoding.

Among methods that have been employed for object detection in video, Ngo et. al. [12] described object detection based on motion and color features using histogram analysis. This technique could process less than 2 frames in one second. Unfortunately, many of the other techniques presented such as [14] did not provide evaluation of time performance. However, it depends on even more involved image processing methods, such as an active contour model, which puts considerable effort into determining the boundary of a shape, and is thus likely to be slower. More recently some compressed domain techniques have been suggested by Wang et al [13]. This system achieved about 0.5 sec/frame for the CIF size on a Pentium III 450 MHz.

Virtually no analysis of techniques that combine eye gaze tracking and scene analysis exist.

1.2. Perceptual Transcoding

This section presents a brief description of our initial eye-tracker based system.

The *Percept Media Transcoder* (PMT) architecture has been designed so that media specific perceptual transcoding modules can be plugged into it without requiring the reorganization of the overall media distribution .

A critical consideration for a real-time perceptual feedback based media transcoding scheme is the *feedback delay.* Feedback delay is the period of time between the instance when the eye position is detected, and when the perceptually encoded frame is displayed. Such delay originates primarily from the network during data transmission and also from the heavy computational complexity of any practical video encoding system. It is important to note that feedback delay can be large and it is also dynamically varying. This dilemma cannot be ignored. As we will show later, feedback delay provides a significant impact in perceptual video compression.

Consequently PMT uses an approach that can operate with dynamically varying feedback delay. Instead of relying only on the human eye acuity matching model, PMT uses the integrated approach of *gaze proximity prediction and containment*. It determines a gaze proximity zone or a Reflex Window. Its goal is to ensure that the bulk of the eye gazes will remain within a certain area with a statistical guarantee given some value of feedback delay.

Note that potential gain in video compression depends on the size of the high quality area on the video frame. Naturally, our design goal was to reduce the size of high quality area without sacrificing the gaze containment, which determines the size of the high resolution area. Our hybrid scheme implemented in PMT combines the Reflex Window with an internal object detection mechanism to reduce the size of the area requiring high quality encoding.

The Reflex Window, Object Window, and also resulting Perceptual Object Window constructions are done in real time.

2. REFLEX WINDOW

2.1. Human Visual Dynamics

Intricate types of eye movements scientists have identified are drift, saccade, fixation, smooth pursuit eye-movement, involuntary saccade. Among them, the following two play the most important roles in the design of the proposed system. (i) Saccades are simultaneous and identical rapid rotations of the eyes that occur between two points of fixations. (ii) Fixations are eye movements that take place when the object of perception is stationary relative to the observer's head: small involuntary saccades, drift, and tremor.

2.2. Reflex Window

The objective of the Reflex Window (W^{RW}) is to contain the fixations by estimating the probable maximum possible eye velocity due to saccades. Given a set of past eye-positions, the W^{RW} represents a zone where the eye will be at during a certain point in the future from its current position with certain likelihood. The acceleration, rotation and de-acceleration involved in ballistic saccades are guided by the muscle dynamics and demonstrate stable behavior. The latency, vector direction of the gaze, and the fixation duration, has been found to be highly dependent on the content, and hard to predict. Therefore we model W^{RW} as an ellipse which is centered at the last known sample location, allowing the gaze to take any direction within the acceleration constraints. If (x_c,y_c) is the current detected eye-gaze position, then we model W^{RW} as an ellipse with center at (x_c,y_c) with half axis $x_R = T_d V_x(t)$ and $y_R = T_d V_y(t)$. See Fig. 2.2.1. T_d is a feedback delay $V_x(t)$ and $V_y(t)$ are the *containment assured eye velocities* (CAV). CAV represents a predicted eye velocity, which will allow for the containment of the targeted amount of eye gazes given a value of feedback delay. The length of the T_d consists of the delay introduced by the network and eye tracking equipment plus the time it takes to encode a particular video frame.

2.3. Eye Velocity Prediction

Future eye gaze position predictions are based on the past positional variances. Our algorithm estimates the past eye velocity components to be used for creating Reflex Window ellipse for a given prediction accuracy goal. We use the following k-percentile algorithm to determine this.

Suppose there are n eye samples during t-th frame. Each eye sample $S(t_i)$ has (x_i,y_i) position on the frame F(t) (in units of pixels). The horizontal and vertical components of the eye velocity are then estimated for each frame as:

$$\hat{V}_x(t) = \sum_{i=1}^{n-1} |x(t_{i+1} - T_d) - x(t_i - T_d)| \tag{2.3.1}$$

$$\hat{V}_y(t) = \sum_{i=1}^{n-1} |y(t_{i+1} - T_d) - y(t_i - T_d)| \tag{2.3.2}$$

Here "n" is the number of samples on the particular frame. "n" can vary per frame. Notation $x(t_i-T_d)$ and $y(t_i-T_d)$ means that eye-samples that system received for frame F(t) are T_d msec late. Thus delayed eye-gazes are represented by coordinates: $x(t_i-T_d)$ and $y(t_i-T_d)$, where $1 \le i \le n$, and n is number of eye-gazes detected by eye-tracker while encoding frame F(t). Real eye-gazes coordinates are detected while encoding frame $F(t+T_d)$. They would have coordinates $x(t_i+T_d)$ and $x(t_i+T_d)$ respectively. The equations for RW center are: $x(t_n-T_d)$ and $y(t_n-T_d)$. In real implementation the center of RW is placed on the last available eye-gaze. Fig. 2.3.1 presents the concept of different types of the eye gazes.

$\hat{V}_x(t)$ and $\hat{V}_y(t)$ are *running average eye velocity* samples (RAV). Let, $\hat{\eta}$ to be *target containment factor* (the percentage of future eye gazes to be contained in RW). To determine CAV we first construct histograms of the past k RAV samples in horizontal and vertical dimensions. Then we determine the $\hat{\eta}$ percentile velocity boundaries within the k samples. These percentile boundaries define the value for CAV. The model considers past "k" RAV samples so that it encompasses at least one eye sample from saccade latency, acceleration, de-acceleration, and fixation or pursuit within that period. Detailed description of CAV calculation is available on our website [8]. See the Fig. 2.3.2 and Fig. 2.3.3 for the calculated RAV and CAV for "Video 1".

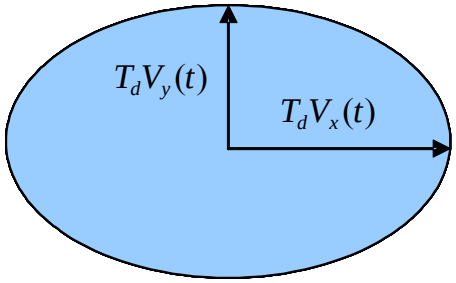

Fig. 2.2.1 Reflex Window diagram

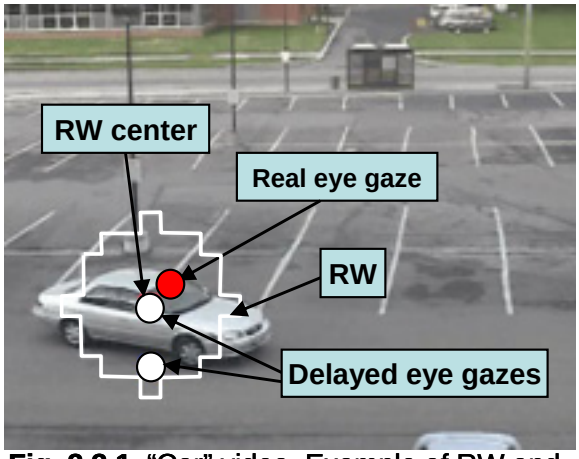

Fig. 2.3.1 "Car" video. Example of RW and different types of eye-gazes.

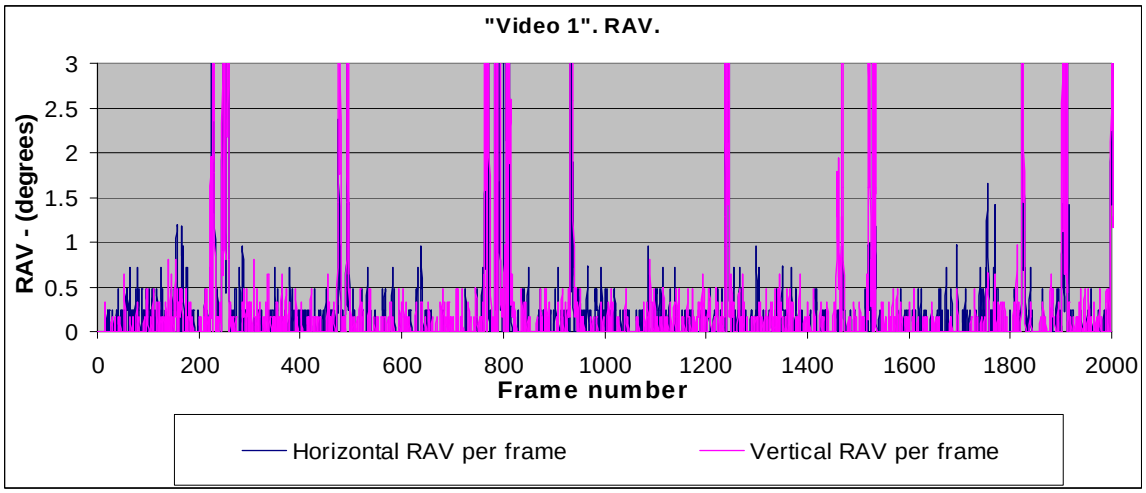

Fig. 2.3.2 Horizontal and vertical running average eye velocity per frame.

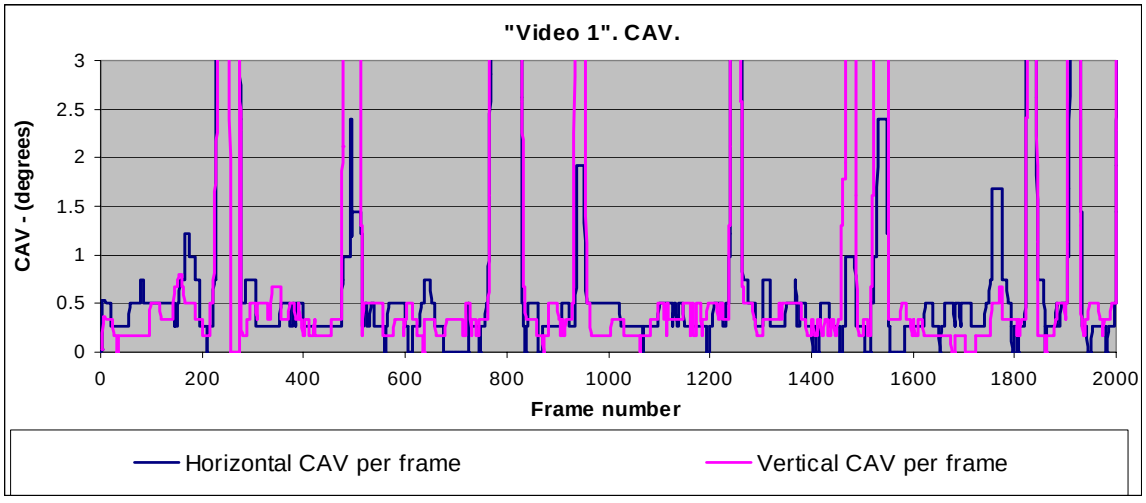

Fig. 2.3.3 Horizontal and vertical containment assured eye velocity per frame. Feedback delay is 1 sec. Each CAV value is calculated based on 20 RAV samples.

3. OBJECT WINDOW

Object Window (W^{TW}) represents the part of the image containing an object on the video frame. The position and the size of such area depend on the location and the dimensions of the object, its speed, etc. The objective of the W^{TW} construction mechanism is to identify and track the object in real time, providing the boundaries of the object's shape. The exact algorithm for W^{TW} construction and implementation is described in [5].

4. HYBRID VISUAL WINDOW

Both W^{RW} and W^{TW} are effective tools for enhanced perceptual video compression. Each of them is based on the different construction method: W^{RW} is based on the eye movement prediction and W^{TW} is based on the video scene analysis. We have thought of possible hybrid scheme which takes both W^{RW} and W^{TW} into consideration to select the area of the video frame which requires high quality encoding. Five models have come into our consideration. Two of them are improvements of Object Tracking Window and three of them are the hybrid windows or Perceptual Object Windows (W^{POW}) created with the help of both W^{RW} and W^{TW}.

4.1. Rectilinear Approximation

When we look at an object, we also look at the area surrounding it. Therefore, we decided to create an approximation around the object boundaries. For the rectilinear approximation of W^{TW} (W^{RTW}) all coordinates of the macroblocks (MB) in W^{TW} are sorted according to their values. W^{RTW} is constructed using the min max values of those coordinates. W^{RTW}= $\{(x_{min}, y_{max}), (x_{min}, y_{min}), (x_{max}, y_{max}), (x_{max}, y_{min})\}$, where $x_{max} = \max\{x_i\}$, $x_{min}=\min\{x_i\}$, $y_{max}=\max\{y_i\}$, $y_{max}=\max\{y_i\}$, where x_i and y_i are MB coordinates and $x_i, y_i \in W^{TW}$. One instance of rectilinear approximation is shown in Fig. 4.1.1.

4.2. Circular Approximation

Another form of W^{TW} approximation is a circular approximation W^{CTW} shown in Fig. 4.2.1. Let (x_i, y_i) represent the coordinates of a macroblock on the video frame. The distance between two macroblocks calculated as: $D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_i)^2}$. Let D_{km} be the maximum possible distance between a pair of macroblocks in W^{TW} $D_{km}=\max\{D_{ij}\}$. Suppose $MB_k[x_k,y_k]$ and $MB_m[x_m,y_m]$ are such macroblocks, then $D_{km} = \sqrt{(x_k - x_m)^2 + (y_k - y_m)^2}$. W^{CTW} is defined as a circle with radius $R=0.5D_{km}$, and center at ($x_c = \frac{x_k + x_m}{2}, y_c = \frac{y_k + y_m}{2}$).

4.3. Hybridization Method A

(i) The idea behind this method is to monitor the center of W^{RW} and watch if it falls inside the boundary of W^{CTW}. When it happens the resulting Perceptual Object Window (W^{POW}) is equal to the intersection of W^{CTW} and W^{RW}. (ii) W^{POW} is equal to W^{CTW} in the case when W^{RW} fully contains W^{CTW}. (iii) In all other cases W^{POW} is equal to W^{RW}. The idea is represented in Fig. 4.3.1.

Assuming that we have a set of macroblocks representing W^{TW} and $MB_R[x_R,y_R]$ is W^{RW} center. Thus in this method:

$$W^{POW} = \begin{cases} \text{i) } W^{CTW} \cap W^{RW} \text{ if } MB_R[x_R, y_R] \in W^{CTW} \\ \text{ii) } W^{CTW} \text{ if } \forall i : MB_i[x_i, y_i] \in W^{CTW} \text{ and } MB_i[x_i, y_i] \in W^{RW} \\ \text{iii) } W^{RW} \text{ in all other cases} \end{cases} \quad (4.3.1)$$

4.4. Hybridization Method B

"Method B" has the same idea behind it as "Method A". Additionally, method B takes into consideration the relative position of W^{CTW} in respect to W^{RW}. In a case when W^{RW}'s center is not contained inside of W^{CTW}'s, "Method B" creates W^{POW} as a half of the W^{RW} directed towards W^{CTW}. The idea of this presented in the Fig. 4.4.1

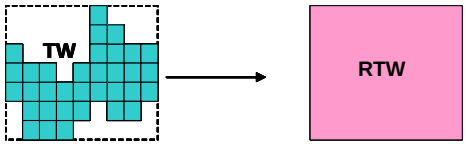

Fig. 4.1.1 Rectilinear Approximation of the Tracking Window.

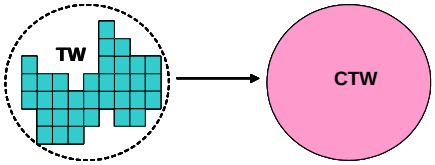

Fig. 4.2.1 Circular Approximation of the Tracking Window.

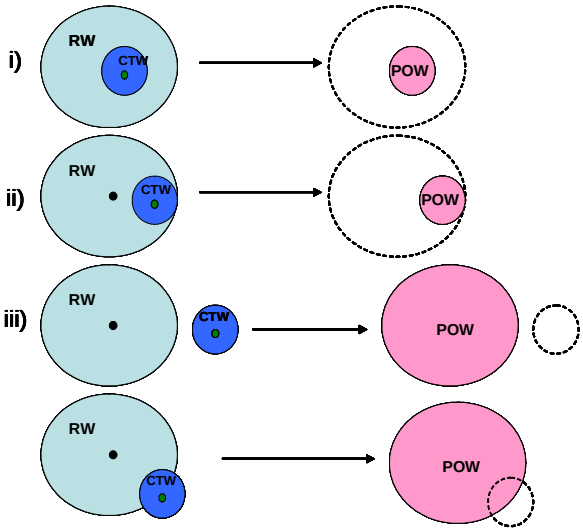

Fig. 4.3.1 Method A diagram.

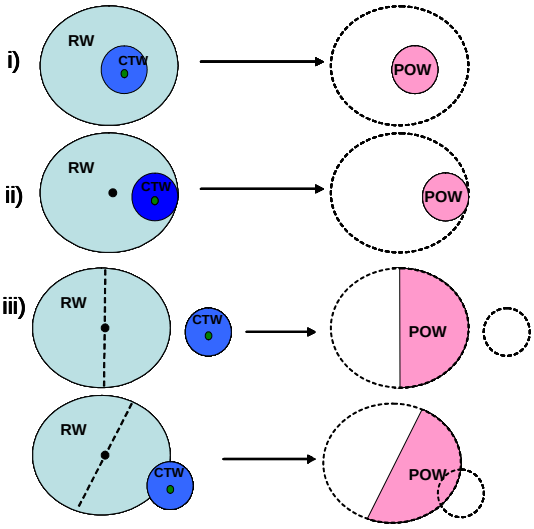

Fig. 4.4.1 Method B diagram.

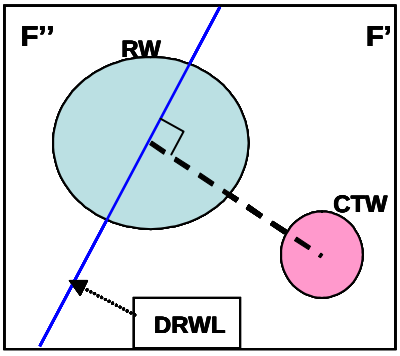

Fig. 4.4.2 DRWL construction diagram.

.

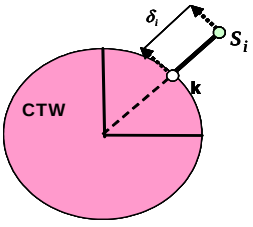

Fig. 4.5.1. Deviation representation.

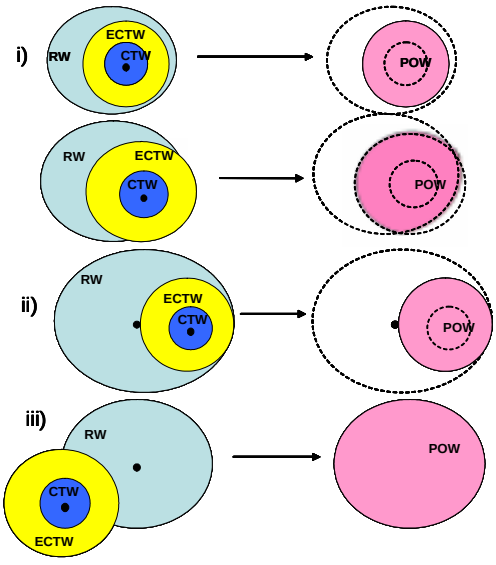

Fig. 4.5.2. Method C diagram.

W^{DRW} is Divided Reflex Window. It is constructed from W^{RW} by splitting W^{RW} in half by Divided Reflex Window Line (DRWL), which is orthogonal to the line going trough W^{RW} center (X_{RW}, Y_{RW}) and W^{CTW} center (X_{CTW}, Y_{CTW}). See Fig-4.4.2. DRWL divides video frame F in two planes F' and F''. F' is the plane which contains the W^{CTW} center - $(X_{CRW}, Y_{CRW}) \in F'$. W^{DRW} is created by the intersection of W^{RW} and F'. $W^{DRW} = W^{RW} \cap F'$.

"Method's B" W^{POW} defined as:

$$W^{POW} = \begin{cases} \text{i) } W^{CTW} \cap W^{RW} \text{ if } MB_R[x_R, y_R] \in W^{CTW} \\ \text{ii) } W^{CTW} \text{ if } \forall i : MB_i[x_i, y_i] \in W^{CTW} \text{ and } MB_i[x_i, y_i] \in W^{RW} \\ \text{iii) } W^{DRW} \text{ in all other cases} \end{cases} \quad (4.4.1)$$

4.5. Hybridization Method C

First we should introduce W^{ECTW} – enhanced W^{CTW}, which is created from W^{CTW} by increasing the radius R of W^{CTW} by some number ε. Quantity ε is adjusted during video playback to provide better performance. In order to choose the best value for ε our algorithm measures how far eye-gazes fall from the boundary of W^{CTW}. Value δ_i - deviation is defined as the distance between W^{CTW} boundary and the eye gaze point S_i. Deviation is calculated only for those S_i located outside of W^{CTW} boundary. See Fig. 4.5.1. Deviation values are collected over some period of time, usually not exceeding the duration of m video frames. The deviation values are processed and new value for ε is chosen for each video frame based on some percentile parameter ϖ. Values of m and ϖ are chosen based on some statistical analysis. They are feedback and video content dependent. For this particular experiment m=10, and $\varpi = 0.7$. With a new radius the W^{RW} center falls into the boundaries of W^{ECTW} much more often. (i) To create final W^{POW} our algorithm chooses the intersection of W^{ECTW} and W^{RW} if W^{RW} center lies inside of W^{ECTW}. (ii) W^{POW} is equal to W^{ECTW} in the case when W^{RW} fully contains W^{CTW}. (iii) In all other cases W^{POW}=W^{RW}. See Fig. 4.5.2.

$$W^{POW} = \begin{cases} \text{i) } W^{ECTW} \cap W^{RW} \text{ if } MB_R[x_R, y_R] \in W^{ECTW} \\ \text{ii) } W^{ECTW} \text{ if } \forall i : MB_i[x_i, y_i] \in W^{ECTW} \text{ and } MB_i[x_i, y_i] \in W^{RW} \\ \text{iii) } W^{RW} \text{ in all other cases} \end{cases} \quad (4.5.1)$$

5. EXPERIMENT

5.1. Setup

We have implemented our system with Applied Science Laboratories high speed eye tracker model 501. The camera that captured the eye position had a sampling rate of 120 samples per second. For this experiment we defined fixation when the eye did not move more than 1 degree in 100 msec.

5.2. Test Data

We have selected three video sequences to test the performance of the proposed hybrid methods.
"Video 1" contained a car driving in a parking lot. The object speed was smooth and continuous.
"Video 2" had two radio controlled toy cars moving at varying speeds. Both toy cars had rapid unpredictable movements. In this video we asked the subject to concentrate on just one car.
"Video 3" had two relatively close up toy cars at offering a large area of attention. Cars moved in different directions inconsistently. Subject was asked to concentrate on only one car.

Each video was MPEG-2 encoded with the original bit-rate of 10MB/s and frame rate of 30fps. Each video clip was around 1 minute long.

The subject who viewed the test videos was familiarized with them before the experiment.

Fig. 5.2.1 and Fig. 5.2.2 show sample frames 233 and 1343 of "Video 1". Fig. 5.2.3 and Fig. 5.2.4 show both W^{RW} (for 90% eye containment) and the W^{TW} as estimated by the eye-tracker only and scene analysis only methods. These also show the actual eye gazes samples on these frames. We see that W^{RW} was able to contain three of the four gaze samples on the frame 233. On the other hand, the W^{TW} failed to contain any. Also note the large coverage of the W^{RW}s. The

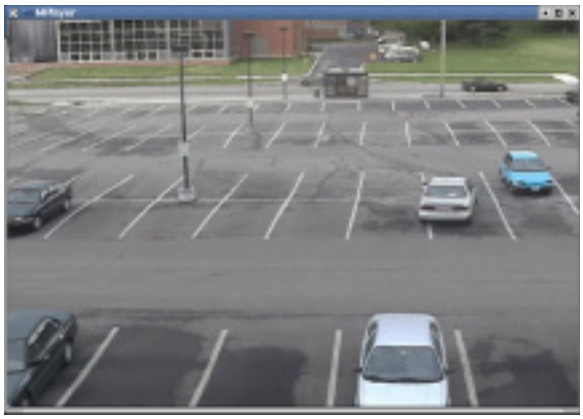

Fig. 5.2.1 Video 1. Frame number 233. Original MPEG-2 encoding scheme. Bit rate 10Mb/s.

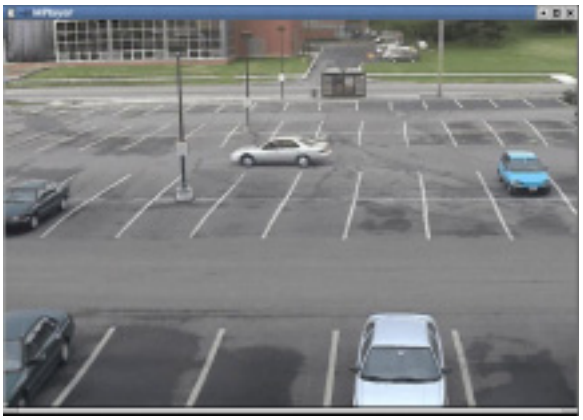

Fig. 5.2.2 Video 1. Frame number 1343. Original MPEG-2 encoding scheme. Bit rate 10Mb/s.

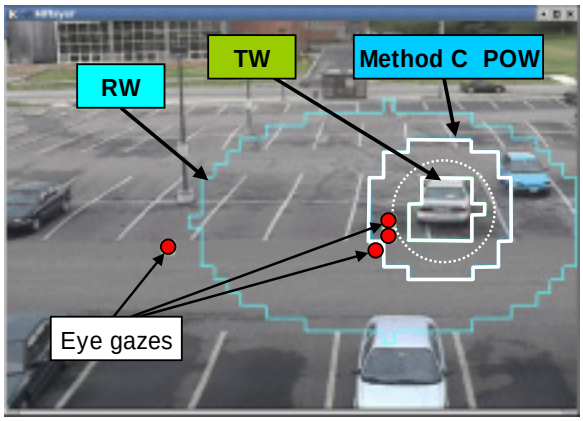

Fig. 5.2.3 Video 1. Frame number 233. RW – Reflex Window. TW – Tracking Window. Method C POW – Perceptual Object Window created using method C, case i). This picture also represents the case of multiple eye gazes on a single video frame, where none of the gazes fall into TW, some of the fall inside POW, some of the fall inside RW, and some of the eye gazes fall outside RW.

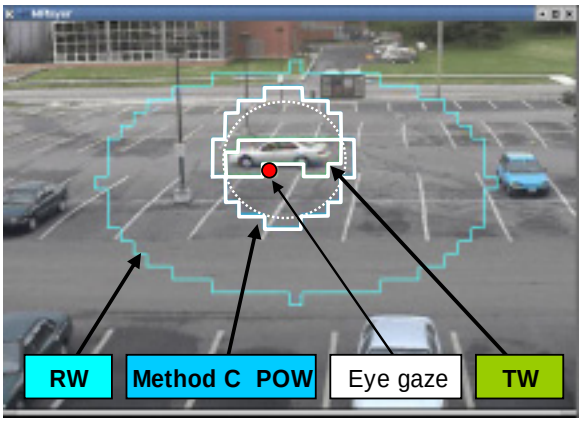

Fig. 5.2.4 Video 1. Frame number 1343. RW – Reflex Window. TW – Tracking Window. Method C POW – Perceptual Object Window created using method C, case i). The eye gaze on this frame falls outside of TW, but inside of POW.

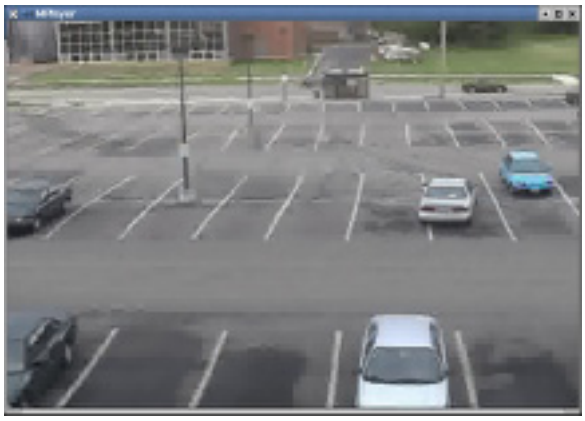

Fig. 5.2.5 Video 1. Frame number 233. Perceptually encoded. Target bit rate 1Mb/s.

Fig. 5.2.6 Video 1. Frame number 233. Perceptually encoded. Target bit rate 1Mb/s.

"Method C" W^{POW} reduced frame coverage and simultaneously was able to contain two of the gazes, though slightly missed one. For frame 1343, the "Method 3" W^{POW} fully contained the gaze. Fig. 5.2.5 and Fig. 5.2.6 provide sample of perceptually encoded frames based of the Method-C W^{POW}. Note that the bit-rate was reduced about 10 times to 1 Mbps. In this bit reduction scheme full resolution was maintained at the W^{POW} macro-blocks. The MPEG-2 TM-5 rate control was used to determine the quantization of the remaining blocks. The actual perceptually encoded video samples, including the originals can be obtained for direct visual appreciation from [16].

6. PERFORMANCE ANALYSIS

To measure the effectiveness of our algorithm we have defined the following two parameters: eye gaze containment and perceptual window coverage efficiency. Perceptual window represents a part of the image requiring high resolution coding. In our case, perceptual windows are W^{TW}, W^{RW}, W^{RTW}, W^{CTW}, $W^{ECTW,}$ and W^{POW} created by each method.

6.1. Eye Gaze Containment

The primary goal of the perceptual encoding is to contain eye fixations within the perceptually encoded window. Ideally, if all gazes are within such a window, it is then possible to design an optimum perceptual encoder. Thus, we defined the quantity *gaze containment* as the fraction of gazes successfully contained within a window:

$$\xi = \frac{\left|E^{w}(t)\right|}{\left|E(t)\right|} \tag{6.1.1}$$

Where, E(t) is the entire eye-gaze sample set. $E^{W}(t) \subseteq E(t)$ is the eye-gaze sample subset contained within an arbitrary window W(t).

6.2. Perceptual Coverage

The other important design goal is to reduce false eye gaze containment. With a large perceptual window more gazes can be contained, however, there will not be any perceptual redundancy to extract. Therefore, we have defined a second performance parameter called *perceptual coverage* for obtaining video frame coverage efficiency by a perceptual window. If F(t) is the size of the viewing frame, and W(t) is perceptual window, then the perceptual coverage is given by (delta for area or volume):

$$\chi(t) = \frac{\left|\Delta(W(t) \cap F(t))\right|}{\left|\Delta(F(t))\right|} \tag{6.2.1}$$

Now we present the performance of each method with respect to these two parameters.

6.3. Analysis of Results

Performance of the Eye-Gazed based System: Figures 6.3.1-6.3.3 provides the results. The left y-axis and the bar-graphs show the perceptual coverage efficiency of each method. The right y-axis and the line curves show the corresponding gaze containment. The leftmost TW (W^{TW}) and rightmost RW (W^{RW}) cases respectively show the performance of the strictly object based method, and strictly eye-gaze based method. In the absence of significant feedback delay, (155 ms or 5 video frames), the eye-tracker based methods offered approximately a 3% frame coverage and roughly a 90% gaze containment. However, when the feedback delay was about 1 second (30 frames), the Reflex Window became quite large, close to 28%. With larger frame coverage there is lesser scope of compression.

Performance of the Pure Object-based System: Now let us look at the case of the pure scene analysis based perceptual encoding attempt. We can see that the advantage of W^{TW} is its smaller coverage area (about 5%). A small coverage of the area of interest creates a potential for high compression. Conversely, its weakness is accuracy of the fovea. As it can be noted, despite the small coverage, W^{TW} actually misses a significant amount of the eye-gazes. Its containment is only about 50%. Thus, a perceptual compression based on just object detection is expected to lack high perceptual quality.

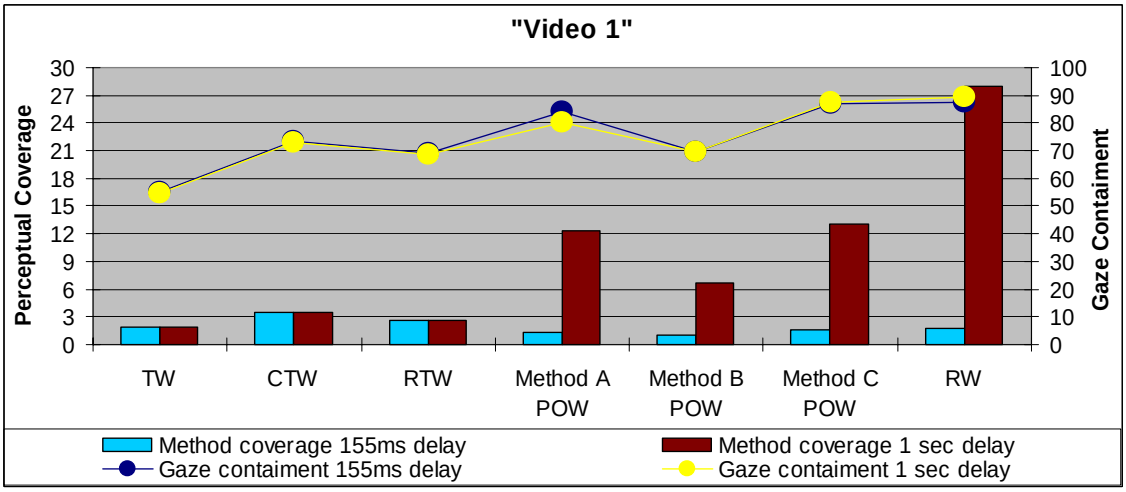

Fig. 6.3.1. Video 1. Gaze containment results for different POW build methods.

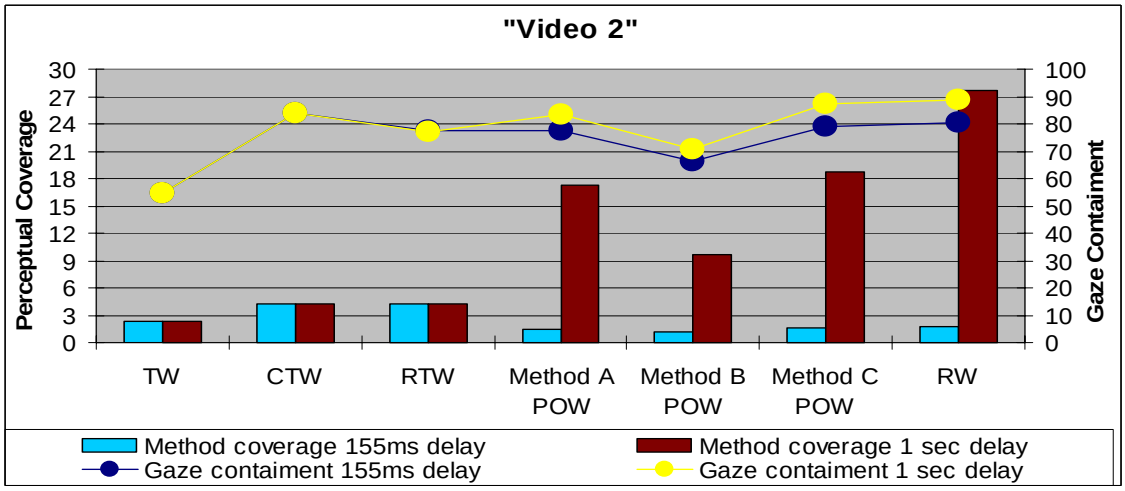

Fig. 6.3.2. Video 2. Gaze containment results for different POW build methods.

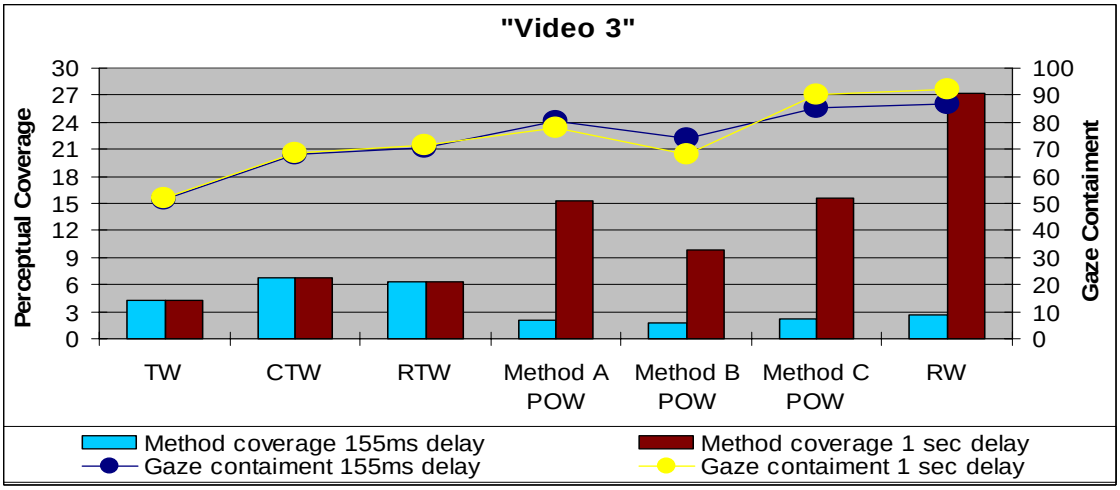

Fig. 6.3.3. Video 3. Gaze containment results for different POW build methods.

Improvement due to Approximations: Before we move to the hybrid techniques, we also present the performances of the two approximations performed based on the pure object approach. The plots CTW (W^{CTW}) and RTW (W^{RTW}) respectively provide corresponding performances. Compared to strict object boundary based TW (W^{TW}), these approximations double the coverage area from 3% to 6%. However, at the same time these improve the gaze containment significantly from 50% to about 70%.

Hybrid Methods: The incorporation of the scene analysis kept the containment near the level of the eye tracking only method (RW), but drastically reduced perceptual coverage. For 1 second feedback delay, "Method A" kept gaze containment near 80%, but the coverage was drastically reduced from 27% to about 9-15%. Among the methods used, "Method B" was more conservative on the side of reducing perceptual coverage. It offered coverage of about 9% with gaze containment of about 70-75%. "Method C" on other hand offered containment almost in the level of pure eye-tracker based method (RW) but reduced the coverage to a level of 15%. The hybrid methods, particularly "Method C", were able to reduce the perceptual coverage from 27% to about 15%, without any significant loss of the eye gaze containment.

Impact of Feedback Delay: Feedback delay is a major performance factor in the proposed hybrid scheme. The longer the delay, the larger was the size of the constructed perceptual window. In the case of a 1 sec delay, the size of the perceptual object window was around 9-15% of the video frame. The instance of a 155 ms delay the perceptual coverage went down to 2-3%. But in each feedback delay scenario hybrid methods reduced the perceptual coverage significantly compared to simply object-based or eye gaze-based compression methods.

Impact of Object Size & Quantity: The hybrid approach has the ability to reduce the size of the perceptual window, making it even smaller than the size of the object itself. Any exclusively eye-tracker methods must use a larger perceptual window due to inherent feedback delay. This is evident in the experiment with the second video in which featured two objects. A scene-only analysis faces ambiguity, as it does not know precisely at which object a person is looking at. In the hybrid method the eye gaze analysis helps in resolving this ambiguity. As seen in Fig-6.3.3, with the 155 ms delay experiment, the coverage of the hybrid method was about 2% compared to about 3% of W^{TW}. Even for one large object the hybrid technique can help in focusing in a smaller area. This is evident in the "video-3" experiment. Here the object window was about 5% (TW), but the hybrid perceptual window covered only about 2% of the frame and was still able to contain 80-90% of the gazes.

7. CONCLUSIONS & CURRENT WORK

Eye trace-based media coding is a promising area. Nevertheless, a number of formidable technical challenges still remain before all characteristics of the human eye can be exploited to engineering advantages. In this paper we have addressed the mechanism of how the eye gaze fovea can be further narrowed in a dynamic environment with augmentation from low grade scene analysis. Opportunities exist for drastic reduction of coverage without any loss of eye-gaze containment. It is important to note that in live video transcoding processing speed is a critical consideration. Consequently, for both scene analysis and eye movement prediction we have used computationally low cost approximation approaches. There are intricate schemes known for the detection of objects in videos. Though these require massive image processing and we could not use them for this scenario.

It is also interesting to note that with few exceptions, most of the previous studies in eye-tracker based media compression have focused on the study of the foveal window degradation around the point of an eye gaze. Even when some type of fovea region was considered, it was of fixed size and static. In this paper, we have focused on a scenario when the perceptual video encoding scheme is affected by a significant feedback delay. This makes the dynamic estimation and optimization of the area that requires the highest quality coding more important than precise calculation of the peripheral acuity degradation.

Further research should be performed to understand the media dependent degradation and coding models when the perceptual window is of dynamic nature.

REFERENCES:

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

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

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

[4] 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.

[5] Javed I. Khan and Zhong Guo, Flock-of-Bird Algorithm for Fast Motion Based Object Tracking and Transcoding in Video Streaming, The 13th IEEE International Packet Video Workshop 2003, Nantes, France, April 2003.

[6] 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.

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

[8] Javed I. Khan, Oleg Komogortsev, "Dynamic Gaze Span Window based Foveation for Perceptual Media Streaming", Technical Report TR2002-11-01, Kent State University, [available at URL http://oahu.medianet.kent.edu/technicalreports.html, also mirrored at http://bristi.facnet.mcs.kent.edu/medianet] November, 2002.

[9] 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.

[10] 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.

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

[12] Ngo, Chong-Wah, Ting-Chuen Pong and Hong-Jiang Zhang, "On clustering and retrieval of video shots", ACM Multimedia 2001, Oct., 2001. pp51-60.

[13] Wang R., H.J. Zhang and Y.Q. Zhang. A confidence measure based moving object extraction system built for compressed domain. 2000.

[14] Kuehne, Gerald, Stephan Richter and Mark Beier, "Motion-based segmentation and contour-based classification of veideo objects", ACM Multimedia 2001, Oct., 2001. pp41-50

[15] Duchowski, A.T., McCormick, Bruce H., "Preattentive considerations for gaze-contingent image processing" in Human Vision, Visual Processing, and Digital Display VI, April 1995, SPIE.

[16] Javed I. Khan, Oleg Komogortsev, "Perceptually Encoded Video Set from Dynamic Reflex Windowing", Technical Report TR2002-06-01, Kent State University, [available at URL http://oahu.medianet.kent.edu/technicalreports.html, mirrored at http://bristi.facnet.mcs.kent.edu/medianet].