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 in real-time eye-gaze based perceptual streaming is how to handle the fast nature of human eye-gaze interaction with relatively complex media coding transcoding scheme, and the control loop delay associated with streaming in the network. This delay requires prediction and increases the size of the area requiring high acuity. 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 newest area of high fidelity media coding. It is believed that our eye samples only a part of the visual plane at any given time. The idea is to decide the spatial distribution of bits with close coherence with the perceptually meaningful shapes, objects and actions present in a scene. A number of researchers have confirmed the effectiveness of such a scheme. However, there are challenges in the engineering of this approach. It is extremely computationally difficult to detect perceptual significance in a given scene. Indeed the nature of perception dictates that our previous experience also plays an important role in visualization process. Thus, a scene itself may not have all the information that guides visual sampling. Researchers in vision and eye-tracking have suggested perceptual coding with direct eye-tracker based detection of visual attention. Only about 2 degree in our about 140 degrees vision span has sharp vision. A fascinating body of research exists in vision and psychology geared towards the understanding of human visual perception system. These techniques take feedback from eye and head tracking devices and attempt to obtain the correct eye-gaze information in regard with the visual plane. Based on the acuity distribution characteristics of human eye around the fovea these methods use variable spatial resolution coding for image compression. These techniques however require precise spatio-temporal information about the eye position. In the case of a large network delay, or encoding delay an eye moves away from its detected location by the time this information is received. This fact severely offset the 2 degree acuity advantage.

Recently, we have performed several experiments to develop a hybrid technique which combines direct eye gaze sampling with a video scene content analysis. It is believed that both scene content and an eye movement pattern determine precise area of human attention. Our hybrid scheme uses both these facts to calculate the exact area of the image requiring 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}.

We show that our technique simultaneously 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. The technique is probably one of the first which combines 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 such system. The two base techniques for scene analysis and eye gaze based Reflex Window has been described in [KhGu01] and [KhKo03]. In this paper we present them briefly in next 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 which 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 [DuMB95b, KUGG98, LoMc00, Duch00b]. [GWPJ98] presented pyramid coding and used pointing device to identify focus. [KhYu96a, KhYu96b] 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 of designing a good display system [LoMc00, DuMB98, Daly98]. For example, [Daly98] used live-eye tracker to determine the maximum frequency and spatial sensitivity for HDTV displays with fixed observer distance. [LePB01] discussed how to optimally control the bit-rate for MPEG-4/ H.263 stream for foveated encoding.

Among the methods employed for object detection in video, Ngo et. al. [Ngo01] 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 [KuRi01] did not provide evaluation of time performance. However, it depends on even more involved image processing methods such as active contour model which spends considerable effort to determine the shape boundary, and is thus likely to be slower. More recently some compressed domain techniques have been suggested such as by Wang et al [WaZh00]. This system achieved about 0.5 sec/frame for the CIF size on a Pentium III 450 MHz.

Almost no works exist on the techniques which try to combine both eye gaze tracking and scene analysis.

1.2. Perceptual Transcoding

Before, we present the hybrid scheme in this section we will provide 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 systems networking.

A critical consideration for such 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 but also from the heavy computational complexity of any practical encoding system. It is important to note that feedback delay can be large and it is also dynamically varying. This fact 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 bulk of the eye gazes will remain within 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 is to reduce the size of high quality encoded 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. The Reflex Window and Object Window and also resulting Perceptual Object Window constructions are done in real time.

2. REFLEX WINDOW

2.1. Human Visual Dynamics

Scientists have identified several intricate types of eye movements such as drift, saccade, fixation, smooth pursuit eye-movement, involuntary saccade. Among them, the following two play most important role in the design of the proposed system: (i) Saccades: are identical and simultaneous very rapid rotations of the eyes that occur between two points of fixations, and (ii) Fixations: 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 is to contain the fixations by estimating the probable maximum possible eye velocity due to saccades. Given a set of past eye-positions, the reflex window predicts a zone where the eye will be at a certain point in future from its current position with target 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 unpredictable. Therefore we model the reflex window as an ellipse centered at the last known sample location, allowing the gaze to take any direction within the acceleration constraints. Let *(x_c,y_c)* is the current eye-location. Then we model the Reflex Window as an ellipse with center at *(x_c,y_c)* with half axis $x_R= T_dV_x(t)$ and $y_R= T_dV_y(t)$. 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 to contain targeted amount of eye gazes given a value of feedback. The length of the feedback delay 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

The future eye gaze position prediction is based on the past positional variances. It estimates the right 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) (position in units of pixels). The estimated horizontal and vertical components of the eye velocity are then estimated for each frame as:

$$\hat{V}_x(t)=\sum_{i=1}^{n-1}|x(t_{i+1}-T_d)-x(t_i-T_d)| \qquad (2.6.1)$$

$$\hat{V}_y(t)=\sum_{i=1}^{n-1}|y(t_{i+1}-T_d)-y(t_i-T_d)| \qquad (2.6.2)$$

Here "n" is the number of samples on the particular frame. "n" can vary per frame. Notation $x(t_i\text{-}T_d)$ and $y(t_{i\text{-}}T_d)$ means that eye-samples that system received for frame F(t) are T_d msec late. Thus delayed eye-gazes are represented by coordinates: $x(t_i\text{-}T_d)$ and $y(t_i\text{-}T_d)$, 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 would are: $x(t_n\text{-}T_d)$ and $y(t_n\text{-}T_d)$. 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 [KhKo03]. See the Fig 2.3.2 and Fig 2.3.3 for the calculated RAV and CAV for "Video 1".

3. OBJECT WINDOW

Object Window (W^{TW}) serves as an area containing an object on the video frame. The position and the size of such area depends on the location and the dimensions of the object, it's speed etc. The objective of the Object Window construction mechanism is to maintain as accurate as possible contour for the object at all times. The algorithm for W^{TW} construction and implementation is described in [KhGu01].

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 integrated scheme which takes both W^{RW} and W^{TW} into consideration to select the area of the video frame which requires high quality encoding. We have considered five models. Two of them are improvements of Object Tracking Window and three of them are the hybrid windows, created with the help of both W^{RW} and W^{TW}. We will call such combined window Perceptual Object Window (W^{POW}).

4.1. Rectilinear Approximation

Generally looks human at both the object of interest and the area surrounding it. Therefore, we decided to create an approximation around the object boundaries. For the rectilinear approximation 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. 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 macroblocks 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, 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:

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

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 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

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:

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

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 that our algorithm measures how far eye-gazes fall from the boundary of W^{CTW}. Value δ_i- deviation 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}.

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

5. EXPERIMENT

5.1. Setup

We have implemented the system with integrated Applied Science Laboratories High speed Eye tracker Model 501. The eye position video capturing camera had rate of 120 samples per second. For this experiment we defined fixation when the eye did not move more than 1 degree in 100msec. The Percept Media Transcoder was modified in a way that it could generate the Reflex Window, track the object and use proposed methods to create Perceptual Object Window in real time.

5.2. Test Data

The description of the test videos as follows:

"Video 1" contained car driving in the parking lot. The object speed was smooth and continuous.

"Video 2" had two radio controlled toy cars moving at the different speeds. Both toy cars had rapid unpredictable movements. In this video we asked subject to concentrate on just one car.

"Video 3" had two relatively close up toy cars at offering a large area of focus. Cars moved in different directions inconsistently. Subject was asked to concentrate only on 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 was viewing the test videos was familiarized with them before the experiment. During the experiment the subject was asked to look at the moving object on the scene.

Fig-5.2.1 and Fig-5.2.2 show sample frames 233 and 1343 of "Video 1". The video had original encoding at 10 Mbps. 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 techniques. These also show the actual eye gazes samples on these frames. As we can see the 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 "Method C" W^{POW} reduced frame coverage and at the same time was able to contain two of the gazes, and 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}. As it can be noted 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 [KhKo02].

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 can be any window which requires 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 window then it is possible to design an optimum perceptual encoder. Thus, we defined the quantity *gaze containment* as the fraction of gazes successfully contained within a window:

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

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. 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):

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

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 provide 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 about 3% frame coverage and about 90% gaze containment. However, when 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 lets 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. However, 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.

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 eye-trace 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 brought down the coverage to a level of 15%. In general 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: Also, it is noticeable that a major performance factor for proposed hybrid scheme is the feedback delay. The bigger the delay the larger is the size of the constructed perceptual window. In the case of 1 sec delay the size of the perceptual object window is around 9-15% of the video frame. In case of the 155 ms delay the perceptual coverage goes down to 2-3%. But in each feedback delay scenario hybrid methods reduce the perceptual coverage significantly comparing to just 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 encoded window making it even smaller than the size of the object itself. Any eye-tracker only methods have to use larger perceptual window due to inherent feedback delay. This is evident in the experiment with the second video. As can be noted this video has two objects. A scene only analysis faces ambiguity because it does not know exactly at what object a person is looking at. In the hybrid method the eye gaze analysis helps in resolving the ambiguity. As it can be seen in Fig-6.3.3, with 155 ms delay experiment, the coverage of the hybrid method is 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 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. However, a number of formidable technical challenges still remain before such characteristics of human eye can be exploited for engineering advantage. 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. It seems 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. Therefore for both scene analysis and prediction we have used computationally low cost approximation approach. There are much involved schemes known for object detection in video. However, these require massive image processing and we could not use them for this scenario.

Also it is interesting to note that most of the previous studies (with few exceptions), in eye-tracker based media compression have focused on the study of the foveal window degradation around the point gaze. Even when some type of fovea region was considered these were fixed sized and static. In this paper, we have focused on the dynamic optimization of the fovea region. It seems the impact of feedback delay involved in any practical system makes dynamic estimation of this fovea region more important that the impact of 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:

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

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

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

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

[KhGu01] 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.

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

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

[KhKo03] 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.

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

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

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

[Ngo01] Ngo, Chong-Wah, Ting-Chuen Pong and Hong-Jiang Zhang, "On clustering and retrieval of video shots", ACM Multimedia 2001, Oct., 2001. pp51-60.

[WaZh00] Wang R., H.J. Zhang and Y.Q. Zhang. A confidence measure based moving object extraction system built for compressed domain. 2000.

[KuRi01] Kuehne, Gerald, Stephan Richter and Mark Beier, "Motion-based segmentation and contour-based classification of veideo objects", ACM Multimedia 2001, Oct., 2001. pp41-50

[Duch95b] 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.

[KhKo02] 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].