A Review on GCD Updates Detectability, and Optimized Methods for Foveated Image & Video Coding

Farzin Farhadi-Niaki
Department of Systems and Computer Engineering
Carleton University
Ottawa, Canada
ffniaki@sce.carleton.ca

Abstract — This paper reviews a Perceptually Optimized Foveation based Embedded ZeroTree Image Coder (POEFIC) that introduces a perceptual weighting to wavelet coefficients prior to control SPIHT encoding algorithm in order to reach a targeted bit rate with a perceptual quality improvement. The study also, provides a new objective quality metric based on a Psychovisual model that incorporates the properties of the HVS, plays an important role in this POEFIC quality evaluation. The perceptual weights for all wavelet subbands are computed based on 1) foveation masking to remove or diminish significant high frequencies from peripheral regions 2) luminance and Contrast masking, 3) the contrast sensitivity function CSF to attain the perceptual decomposition weighting.

On the other hand, image and video coding is an optimization problem. A successful image and video coding algorithm conveys a good tradeoff between visual quality and other coding performance measures, such as robustness, complexity, compression, scalability, and security. This review pursues two recent styles in image and video coding research. One is to integrate human visual system (HVS) models to progress the current state-of-the-art of image and video coding algorithms by better utilizing the properties of the projected receiver. The other is to design rate scalable image and video codecs, which allow the extraction of coded visual information at continuously altering bit rates from a single compressed bitstream.

Prior to aforesaid studies, this paper reviews the work has been done to answer the question “how late can you update gaze-contingent multisolutional displays without detection?”.

Index Terms—GCD update delay, foveated image, video coding, HVS, wavelet.

I. INTRODUCTION

This study examines perceptual disruptions in gaze-contingent multisolutional displays (GCMRDs) as a result of delays in updating the center of resolution after an eye movement. GCMRDs can be used to save processing resources and transmission bandwidth in many types of single-user display applications, such as simulators, video-telephony, virtual reality, and remote piloting. It has also been found that 60 ms delay of image update after an eye movement did not considerably increase the detectability of image degredation and/or motion transients due to the update. This is good news for designers of GCMRDs, since 60 ms is sufficient time to update many GCMRDs after an eye movement without disrupting perception. Moreover, the study found that longer decades caused greater blur and/or transient detection as the eyes move further into the low-resolution periphery, effectively degrading the image resolution at fixation before the update. In GCMRD applications where longer eyes movements are more likely (e.g., displays with relatively distant objects), this problem could be solved by increasing the size of the area of highest resolution.

The psychovisual researches reveal that spatially, the resolution, or sampling density, has the highest value at the point of the fovea and falls down dramatically from that point as a function of eccentricity. This paper also reviews an optimized foveation based image coding quality (POEFIC) algorithm, which takes advantage of various Psychovisual quality models utilizing the human visual system quality criteria (HVS), to optimize foveated image wavelets coefficients weighting and progress the visual quality of its coded version. An objective metric for foveation based image namely, quality wavelet index metric FWQI yields a quality scale called Foveated probability scale FPS, whose test outcome display very good performance in terms of quality measurement.

It has been envisioned that network visual services, such as videoconferenceing, telemedicine, video-on-demand, and network video broadcasting will become ever-present in the current century. Therefore, network visual communication has become a dynamic research area in recent years. A great challenging problem to implement a video communication system is inefficient available bandwidth of the networks for the delivery of the large amount of the video data. Improvement of video compression techniques is a solution for this problem. The quality of a video codec is application-dependent. Foveated image and video coding is directly related to Region-of-Interest (ROI) image and video coding. The main distinction with respect to conventional ROI processing is that the “interest” is continuously space-variant and conforms to HVS characteristics. A fixed foveation model has been used in most of foveation algorithms. Non-flexibility in “adapting to different foveation depths” and “being implemented in a rate scalable manner”, are some drawbacks of these methods.
II. HOW LATE CAN YOU UPDATE GAZE-CONTINGENT MULTIRESOLUTIONAL DISPLAYS WITHOUT DETECTION?

The following sections review the Ref. [2].

A. Background

Users of virtual reality, teleoperation, video-telephony, simulations, and other single-user applications often need large, high-resolution displays exceeding limits on bandwidth and/or computation resources. One solution for these limitations is to remove/reduce detail that users cannot resolve in the visual periphery.

Gaze-contingent multiresolutional displays (GCMRDs) do just that, by dynamically displaying high-resolution information wherever the user is looking, as indicated by a gaze tracker, and lower-resolution elsewhere. Human factors research on GCMRDs has primarily focused on two key questions:

(1) What are the limits of peripheral visual resolution when viewing scene images?

(2) What are the perception and performance costs associated with reducing image resolution below these limits?

However, far fewer studies have tried to answer another critical question for designers of GCMRDs: How late can you update the center of highest resolution after the user has moved his/her eyes, without disrupting perception or performance? The current study provides an answer to this question. The top panel of Fig. 1 schematically represents the image update process that must occur at the end of each saccade.

While ideally one would update the display immediately at the end of each saccade (i.e., eye movement), this is impossible in practice because it takes time to identify both when a saccade has ended and where the eyes are to render the new multiresolutional image, transmit, and display it. Rendering the image alone can take between 25–150 ms.

Such updating delays could cause perceptual difficulties in two ways. First, when a new fixation begins, prior to the update, the fixated region still has reduced image resolution which may hinder perception. Second, when the update occurs, the change in image resolution may be perceived as a motion transient, which may disrupt perception.

B. The Current Study

The current study directly measures image update delay effects on conscious perception of image degradation, by having viewers detect image blur or motion transients in a GCMRD while factorially varying delay and image filtering levels.

Three levels of image filtering have been used to bracket the updating detection threshold. The filtering E2 values were 6.22°, 3.11°, and 1.55° (refer to Fig. 2).

C. Results

As shown in Fig. 3 and 4, the primary analyses examined the effects of image filtering levels and image updating delays on the detection of image blur and/or motion transients in GCMRDs. Further analyses examined the effects of saccade length on detection rates. In addition, a preliminary analysis examined whether tasks affected detection rates.

Fig. 1. "Top: schematic of a GCMRD over time; bottom: same for a dual delayed occasional GCMRD."[2]

Fig. 2. "A set of three example images for filtering levels E2 = 6.22, 3.11, and 1.55 degrees."[2]

Fig. 3. "Proportion detection of occasionally presented gaze-contingent multiresolutional images as a function of filtering level (control, E2 = 6.22, 3.11, and 1.55 degrees) and update delay (5, 20, 40, 60, and 80 ms)."[2]

Fig. 4. "Proportion detection of occasionally presented gaze-contingent multiresolutional images as a function of filtering level (control, E2 = 6.22, 3.11, and 1.55°) and preceding saccade length trintile (lower, middle, upper). Saccade lengths in degrees for the three trintiles were: lower trintile < 1.43° , 1.43° ≥ middle trintile ≤ 3.32°, and upper trintile > 3.32°."[2]
III. A PERCEPTUALLY OPTIMIZED FOVEATION
BASED WAVELET EMBEDDED ZEROTREE IMAGE
CODING

The following sections review the Ref. [3].

A. Background

The psychovisual experiments demonstrates that spatially,
the resolution, or sampling density and contrast sensitivity
decrease dramatically with increasing the viewing angle
namely called eccentricity with respect of that point of fixa-
tion.

The motivation behind foveation image compression
scheme is that there exists considerable high-frequency infor-
mation redundancy in the peripheral regions, so much more
efficient representation of images can be obtained by remov-
ing or reducing such information redundancy, based on the
foveation point(s) and the viewing distances. The first aim of
that scheme is foveation filtering, which foveate a uniform
resolution image, such that when the human eyes gaze at the
point of fixation, they cannot distinguish between the original
and the foveated versions of that image.

In practice, different methods approximate perfect fovea-
tion filter such as; the pyramid structure is suggested to fove-
ate images; the foveation filter consists of a bank of lowpass
filters having variable cutoff frequencies; the structure of fo-
veation filter based on Laplacian pyramid architecture; or the
proposed wavelet based foveation method applies a nonuni-
form weighting model.

Great success has been obtained recently by a class of
wavelet image coding algorithms oriented region of interest
(ROI), such as the standard JPEG2000 and the Embedded Fo-
vation Image Coding (EFIC) algorithms.

B. Discrete Wavelet Pyramid Decomposition

The coder is a combination of 5 function stages, as shown
in Fig. 5.

C. Foveation Mask Weighting Setup

In this operation it has to locate the foveation point to de-
termine the foveation mask to weight the decomposed image;
as a result all frequencies around the region of interest will be
either reduced or removed from the image spectrum.
In first levels, a great amount of frequencies are removed, but
approximately the whole low frequencies are kept and taken
into account in coding.

As shown in Fig. 6, the observer is progressively unable to
detect high frequencies in image when distance increases.

D. Luminance and Contrast Masking

In this work, three visual phenomena are modeled to com-
pute the perceptual Weighting Model SetUp matrix: the JND
thresholds or Just Noticeable Difference, Luminance Masking
(also known as light adaptation), Contrast Masking and the
Contrast Sensitivity Function CSF. This model correlates well
with the famous cortical decomposition (Human Visual Cortex
field).

E. Embedded Zerotree Wavelet Coding

Embedded ZeroTree wavelet coding is a very effective
and computationally inexpensive technique for image com-
pression. Its principles of computation algorithm are (1) wave-
let pyramid decomposition of the image, (2) partial ordering of
the transform coefficients by the highest bit plane of the mag-
nitude, with the ordering information encoded by means of a
set partitioning algorithm that is reproduced at the decoder, (3)
ordered bit plane transmission of refinement bits, and (4)
exploitation of the self-similarity of the image wavelet pyramid
decomposition across different scales.

F. Quality Measure and Experimental Results Discuss

In order to evaluate or compare image compression tech-
niques we need to reliably measure the quality of coded im-
ages by taking into account the famous observer mean opinion
score (MOS). Recently, techniques based on multiple channel
models of the HVS have been shown to improve correlation with the MOS. From these HVS models it is possible to predict, on a pixel by pixel basis, if the noise introduced in the compressed image will be visible to a human observer. The wavelet transform is one of the most powerful techniques for image compression, because of its similarities to the multiple channel models of the HVS. The DWT decomposes the image into a limited number of spatial frequency channels, with respect to the cortical decomposition.

A wavelet based image quality metric, namely, foveation wavelet Quality Index FQWI predict visible differences between the original and degraded image, which yields a quality measure scale called the Probability Scale PS, plays an important role in the CODEC in terms of image Quality Measurement. This factor means the ability of detecting a distortion in a subband \( (\lambda, \theta) \) at location \((i, j)\) in the DWT field.

\[
P_S = \exp \left( -\sum_{(i,j)}^l \left| \frac{1}{\lambda} \right| \right) \]

The greater this factor is the best the decoded image quality is compared to the original or full reference image. Fig. 8 shows the comparison between POEFIC algorithm and SPIHT algorithm using 8 bits per pixel gray scale images.

Fig. 7. “Perceptual masking model SetUp.”[3]

Fig. 8. “Foveation Wavelet Quality Index FQWI of POEFIC vs SPIHT BOAT image at 0.015625 bpp, 0.0625 bpp and 0.25 bpp.”[3]

Table I

<table>
<thead>
<tr>
<th>Quality Gain (%)</th>
<th>POEFIC vs SPIHT</th>
<th>Test Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>V = 1</td>
<td>7.1936</td>
<td>1.1306</td>
</tr>
<tr>
<td>V = 3</td>
<td>8.8763</td>
<td>7.5951</td>
</tr>
<tr>
<td>V = 6</td>
<td>6.8496</td>
<td>7.2226</td>
</tr>
<tr>
<td>V = 10</td>
<td>4.1526</td>
<td>3.9780</td>
</tr>
</tbody>
</table>

IV. FOEVAITION SCALABLE VIDEO CODING (FSVC) WITH AUTOMATIC FIXATION SELECTION

The following sections review the Ref. [1].

A. Background

Designing a video coding and communication system is a complicated task. The first issue that needs to be considered is the quality-compression performance, which aims to provide the best quality decoded video with the minimal number of bits. Depending on the application, there are many other issues related to the goodness of the video codecs. For example, low computational complexity is usually required for real-time applications. In many cases, parallelizability is a desired feature to improve speed. Satisfying a low memory requirement is desirable in many applications to achieve easy buffering and easier embedded implementations on digital signal processors. As shown in Fig. 10, with a continuously rate scalable codec, the data rate of the video being delivered can exactly match the available bandwidth on the network.

B. Basic Methods and General Framework

a) Foveation point (s) setup

The FSVC system first divides the picture being encoded into blocks with a size of 16x16, and the candidate foveation points are limited to the centers of these blocks. By using this strategy, computation is considerably reduced and only one bit for each block is needed to encode the foveation point selection information.

b) Framework of the encoding system

Similar to many other video coding methods, FSVC first divides the input video sequence into Groups Of Pictures (GOPs). Each GOP has one intra-coding frame (I frame) at the beginning and the rest are predictive coding frames (P frames). The general framework for the encoding of I frames and P frames is given in Fig. 11.

The encoding of the I frame is the same as the EFIC algorithm developed for still image coding. Firstly, apply the DWT and obtain the wavelet coefficients. Secondly, the foveation point selection scheme is applied and the HVS model is calculated to determine the visual importance of the wavelet coefficients. The importance value of each wavelet coefficient

Fig. 9. Zelda image compression results. The images of the left columns are for SPIHT ONLY coded images. The images of the right columns are for the visually optimized SPIHT POEFIC coded images. The bit rates from top to bottom are 0.0625 bpp and 0.25 bpp, at observation distance of \( V = 4 \). [3]

Fig. 10. “Bitstream scaling in rate scalable video communications. Each bar represents the bitstream for one frame in the video sequence.”[11]
Fig. 11. “General framework of the FSVC encoding system.”[1]

is then used to weight the wavelet coefficient. Finally, the modified SPIHT algorithm [1] is employed to generate the embedded bitstream. The encoding of the P frames is more complicated. The idea of using P frames in video coding is to exploit temporal redundancy between adjacent frames in the video sequence. Prediction of the current frame from its previous frame is the key technique to make use of temporal redundancy. Motion Estimation (ME) and Motion Compensation (MC) techniques have been successfully used for this purpose. The DWT is applied to the prediction error frame, and the resulting coefficients are weighted and coded with the embedded encoding algorithm. The HVS modeling techniques are different for I frames and P frames. During the encoding process, a rate control algorithm is used to allocate bits to each frame. The allocation is determined by the available bandwidth, user requirements, the HVS modeling results, and the frame prediction error.

C. Implementation of FSVC

The implementation of FSVC focuses on developing an automated foveation setup approach for video sequences with human faces. Furthermore, an adaptive algorithm is proposed for the prediction of the current frame from motion compensated previous frames.

a) Determination of foveation points

Human faces are probably the most frequently focused regions by human observers. A face-foveated video coding algorithm will be very useful to effectively enhance the visual quality in many specific video communication environments such as videoconferencing.

The methods to select foveation points for I frames and P frames are different. For I frames, it first detects face areas as regions of interest and puts foveation points inside those regions.

For the foveation point selection of P frames, FSVC focuses on the regions in the current P frame that provide us with new information from its previous frame. Usually, the prediction errors in those regions are larger than other regions. Therefore, FSVC mainly selects foveation points in those regions with prediction errors larger than a threshold value. The drawback of this method is that the face regions will lose fixation. To solve this problem, an unequal error thresholding method is used to determine foveation regions in P frames. Therefore, a much smaller prediction error threshold value is applied to capture the changes occurring in the face regions (Fig. 12).

b) Adaptive frame prediction

In fixed rate ME/MC based video coding algorithms, a common choice for frame prediction is to use the feedback decoded previous frame as the reference frame for the prediction of the current frame. With this choice, the prediction frames are exactly the same at the encoder and the decoder. However, this choice is infeasible for continuously rate scalable coding because the decoding bit rate is the choice of the decoder and is unavailable to the encoder. There are several solutions to this problem.

This study proposes a solution to this problem, where the original motion compensated frame and the base bit rate decoded and motion compensated frame are combined to make a prediction. The combination is adaptively changed using the foveation model. The encoder and decoder sides of the new frame prediction algorithms are shown in Figs. 13 and 14, respectively. In Fig. 15, it can be observed that the face region and the relative moving information between frames are captured very well with the automated foveation point selection algorithm.

Fig. 12. “Foveation point selection of the “News” video sequence. (a) I frame foveation point selection; (b) P frame foveation point selection with equal error thresholding; and (c) P frame foveation point selection with unequal error thresholding.”[1]

Fig. 13. “Adaptive frame prediction: encoder side.”[1]

Fig. 14. “Adaptive frame prediction: decoder side.”[1]
with the best quality video he/she can get in terms of foveated quality measurement.

Finally, FSVC is a good choice for interactive video communications, where the users are involved in giving feedback information to the other side of the communication system. The feedback information may be regions or objects of interest and can be converted into knowledge about the video sequence inside the FSVC encoder. Consequently, improved video quality can be achieved.

V. CONCLUSION

The current study has tested the effect of update delays on the detectability of image degradation and/or motion transients in GCMRDs. Update delays had a considerable impact on perception of image decimation. The analysis of the saccade length effects indicated that following longer eye movements, image blur was more evident. This also suggests a three-way interaction between update delay, filtering level, and saccade length.

This study has also reviewed a Perceptually Optimized Embedded Foveation based ZeroTree Image Coder algorithm named POEIFIC, which utilizes a range of Human Visual System HVS model to accomplish the aim of improving the perceptual quality of the reconstructed images versus the quality achieved by the conventional embedded coders particularly standard SPIHT. The provided perceptual model contains Luminance masking, Contrast masking and Contrast Sensitivity function CSF with an optimal implementation.

Furthermore, in this paper, a new wavelet-based scalable foveated video coding system, FSVC, has been reviewed. A foveation-based HVS model, which plays an important role in the system, helps the coding system to foveate on the visually important components in the video sequence. FSVC is a flexible prototype that can integrate various kinds of foveation point selection schemes to fit in different application environments. A novel automated foveation point selection method and an adaptive frame prediction algorithm is provided. By using the adaptive frame prediction algorithm, error transmission is well controlled, while better frame prediction is achieved as well. FSVC is very good for special purpose video communication applications such as telemedicine and video-conferencing, where a lot of prior information is available to the encoder. FSVC is also very suitable for dynamic variable bit rate network video transmission. Moreover, FSVC provides greater flexibility for multiuser and heterogeneous network video communications.

REFERENCES