Foveated video quality assessment

Sanghoon Lee, Marios S. Pattichis, and Alan C. Bovik, *Fellow, IEEE*

Abstract

Most image and video compression algorithms that have been proposed to improve picture quality relative to compression efficiency have either been designed based on objective criteria such as signal-to-noise-ratio (SNR) or have been evaluated, post-design, against competing methods using an objective sample measure. However, existing quantitative design criteria and numerical measurements of image and video quality both fail to adequately capture those attributes deemed important by the human visual system, except, perhaps, at very low error rates. We present a framework for assessing the quality of, and determining the efficiency of *foveated* and compressed images and video streams. Image foveation is a process of nonuniform sampling that accords with the acquisition of visual information at the human retina. Foveated image/video compression algorithms seek to exploit this reduction of sensed information by nonuniformly reducing the resolution of the visual data. We develop unique algorithms for assessing the quality of foveated image/video data using a model of human visual response. We demonstrate these concepts on foveated, compressed video streams using modified (foveated) versions of H.263 that are standard-compliant. We find that quality vs. compression is enhanced considerably by the foveation approach.

The authors are with the Laboratory for Image and Video Engineering (LIVE), Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084 USA.

Corresponding author: Professor Alan C. Bovik, Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084 USA; Phone: (512) 471-5370; Fax: (512) 471-1225; E-mail: bovik@ece.utexas.edu.

I. Introduction

Human visual perception is characterized by a variable resolution across the field of view, with the highest resolution occurring at and near the point of fixation, and decreasing away from this point, as a function of eccentricity, because of the non-uniform distribution of photoreceptors on the retina [1]. The point of fixation is projected onto the *fovea*-the area of densest sampling, and the overall variable resolution data is called a *foveated image*. Hence, if a foveated image is artificially created by removing the undetectable frequencies of an original image (presupposing a point of foveation), then the foveated image will appear the same as the original image. Fig. 1 (a) shows a foveated version of the original image where the foveation point is indicated by "x". In the foveated image, high frequencies away from the foveation point have been removed. If attention is kept focussed at the foveation point, then (depending on the reproduced image size and the viewing distance), the foveated image has the same appearance as the original.

It is possible that the assessment of video quality may be enhanced by taking foveation into account. Certainly, this is an aspect of visual function that has been largely ignored in the design of image quality metrics. This requires the development of objective criteria for measuring foveated image/video quality and against the compression gain afforded by the data reduction. We pursue this goal in this paper by introducing one such criterion: *FSNR*, which is the *foveal signal-to-noise ratio*. We apply this metric to the task of foveated image/video quality assessment.

II. Signal-to-noise ratio in curvilinear coordinates

A. FSNR definition

Consider the coordinate mapping from $\mathbf{x} = (x_1, x_2)$ to $\mathbf{\Phi} = (\Phi_1, \Phi_2)$ given by $\mathbf{\Phi}(\mathbf{x}) = [\Phi_1(x_1, x_2), \Phi_2(x_1, x_2)]$. This coordinate transform defines a one-to-one correspondence between $\mathbf{x}$ and $\mathbf{\Phi}(\mathbf{x})$ under the conditions: Φ_1 and Φ_2 are continuous, and have a single-valued inverse. Then, $\mathbf{\Phi}(\mathbf{x})$ is called "curvilinear coordinates".

For a given 2-D original image $o(\mathbf{x})$, $\mathbf{x} \in R^2$, let its Fourier transform be $O(\mathbf{\Omega})$, $\mathbf{\Omega} \in R^2$. If $O(\mathbf{\Omega}) = 0$ for $|\mathbf{\Omega}| \geq \Omega_o$, then $o(\mathbf{x})$ is an Ω_o-band-limited image and denoted as $o(\mathbf{x}) \in B^{\Omega_o}$ which is the space of 2-D bandlimited signals. Let $o(\mathbf{x})$ be the original image displayed on the monitor, $r(\mathbf{x})$ be the reconstructed (decompressed) image displayed on the monitor, $g(\mathbf{x})$ be the image formed on the human eye, $h(\mathbf{\Phi}(\mathbf{x}))$ be the image of $g(\mathbf{x})$ in the curvilinear coordinates,

$v(\mathbf{x})$ be the foveated image of $o(\mathbf{x})$, and finally, $z(\mathbf{\Phi}(\mathbf{x}))$ be the image of $v(\mathbf{x})$ in the curvilinear coordinates. The relationships between the various images are given by $g(\mathbf{x}) = F_v^c(r(\mathbf{x}))$, $g(\mathbf{x}) = h(\mathbf{\Phi}(\mathbf{x}))$, $v(\mathbf{x}) = F_v^c(o(\mathbf{x}))$, $v(\mathbf{x}) = z(\mathbf{\Phi}(\mathbf{x}))$ where F_v^c denotes the process of $foveationfiltering$ in the continuous spatial domain,

Suppose that a foveated image $v(\mathbf{x})$ with local bandwidth $\Omega_f(\mathbf{x}) \leq \Omega_o$ is derived from the image $o(\mathbf{x})$. In such a case, $B^{\Omega_f(\mathbf{x})}$ becomes the space of locally band-limited signals and is denoted $v(\mathbf{x}) \in B^{\Omega_f(\mathbf{x})}$. Given $\mathbf{\Phi}(\mathbf{x})$, $v(\mathbf{x}) \in B^{\Omega_f(\mathbf{x})}$ is mapped into $z(\mathbf{\Phi}(\mathbf{x})) \in B^{\Omega_c}$ according to $v(\mathbf{x}) = z(\mathbf{\Phi}(\mathbf{x}))$ where B^{Ω_c} is the space of Ω_c-band-limited images. Fig. 1 (a) and (b) show a pair of the foveated images over cartesian coordinates and the curvilinear coordinates respectively.

Once the foveated image is mapped into a uniform image via the coordinate transformation, image/video quality assessment can be accomplished in the same way as with uniform coding (with possible modifications to account for structural distortions). The objective criterion that is used here is the *foveal signal-to-noise ratio*, or FSNR.

Let $S_o \subset \mathcal{R}^2$ be a spatial region of one frame of the original video sequence, and displayed on a monitor over the spatial $\mathbf{x}$ domain, and A_o be the associated area of this region in curvilinear coordinates as shown in Fig. 1. Then, $A_c = \int\limits_{S_o} J_\Phi(\mathbf{x}) d\mathbf{x}$ where $J_\Phi(\mathbf{x})$ is the *jacobian* of the coordinate transformation ($\mathbf{x}$ to $\mathbf{\Phi}(\mathbf{x})$).

Definition 1: M_e^c(mean square error in curvilinear coordinates for continuous images) - Assume that there exists a coordinate transformation $\mathbf{\Phi}$ which maps $v(\mathbf{x})$ and $g(\mathbf{x}) \in B^{\Omega_f(\mathbf{x})}$ into $z(\mathbf{\Phi})$ and $h(\mathbf{\Phi}) \in B^{\Omega_c}$. Then, $M_e^c = \frac{1}{A_c} \int\limits_{S_c} \left[z(\mathbf{\Phi}) - h(\mathbf{\Phi})\right]^2 d\mathbf{\Phi}$. The MSE M_e^c can be also expressed in terms of $v(\mathbf{x})$, $g(\mathbf{x})$ and $J_{\mathbf{\Phi}}(\mathbf{x})$: $M_e^c = \frac{1}{A_c} \int\limits_{S_o} \left[v(\mathbf{x}) - g(\mathbf{x})\right]^2 J_{\mathbf{\Phi}}(\mathbf{x}) d\mathbf{x}$.

The PSNR value is calculated from the discrete images $o(\mathbf{x}_n)$ and $r(\mathbf{x}_n)$. Using a discrete foveation filter $F_v^d(\mathbf{x}_n)$, both $v(\mathbf{x}_n)$ and $g(\mathbf{x}_n)$ can be obtained from $o(\mathbf{x}_n)$ and $r(\mathbf{x}_n)$ respectively. On the other hand, $v(\mathbf{x})$ and $g(\mathbf{x})$ are foveated by the human eye (effectively, by a biological foveation filter). We assume that $v(\mathbf{x}_n)$ and $g(\mathbf{x}_n)$ are the sampled images from $v(\mathbf{x})$ and $g(\mathbf{x})$.

Definition 2: M_e^d (mean square error in curvilinear coordinates for discrete images) - Given a coordinate transformation $\mathbf{\Phi}$ applied to two discrete images, $v(\mathbf{x}_n)$ and $g(\mathbf{x}_n)$, we have $M_e^d = \left[\sum_{n=1}^N \bar{J}_{\mathbf{\Phi}}(\mathbf{x}_n)\right]^{-1} \sum_{n=1}^N \left[v(\mathbf{x}_n) - g(\mathbf{x}_n)\right]^2 \bar{J}_{\mathbf{\Phi}}(\mathbf{x}_n)$ where $\bar{J}_{\mathbf{\Phi}}(\mathbf{x}_n) = \int\limits_{\mathbf{x} \in s_n^o} J_{\mathbf{\Phi}}(\mathbf{x}) d\mathbf{x}$.

Definition 3: FSNR (foveal signal-to-noise ratio) - The objective fidelity criterion FPSNR (foveal peak signal-to-noise ratio) is given by FPSNR $= 10 \log_{10} \left[\max[v(\mathbf{x}_n)]^2 (M_e^d)^{-1}\right]$ and the FSNR(foveal mean square signal-to-noise ratio) is FSNR $= 10 \log_{10} \left[\frac{\sum_{n=1}^N v(\mathbf{x}_n)^2 \bar{J}_{\mathbf{\Phi}}(\mathbf{x}_n)}{\sum_{n=1}^N [v(\mathbf{x}_n) - g(\mathbf{x}_n)]^2 \bar{J}_{\mathbf{\Phi}}(\mathbf{x}_n)}\right]$.

When measuring foveated video quality on a uniform grid, $v(\mathbf{x}_n)$ and $g(\mathbf{x}_n)$ are replaced by $o(\mathbf{x}_n)$ and $r(\mathbf{x}_n)$.

Given a foveated image, let f_{p_n} be the local frequency at the n^{th} point. A sampling matrix $\mathbf{V}_n$ can be found which describes the local frequency and which avoids aliasing if the image is bandlimited. Assume that $\bar{J}_{\mathbf{\Phi}}(\mathbf{x}_n)$ is proportional to the sampling density: $\bar{J}_{\mathbf{\Phi}}(\mathbf{x}_n) = \frac{c_1}{|\det \mathbf{V}_n|} = c_2 f_{p_n}^2$ where c_1 and c_2 are constants. Then, the FPSNR becomes FPSNR $= 10 \log_{10} \frac{\left(\sum_{n=1}^{N} f_{p_n}^2\right) \max[v(\mathbf{x}_n)]^2}{\sum_{n=1}^{N} [v(\mathbf{x}_n) - g(\mathbf{x}_n)]^2 f_{p_n}^2}$.

III. Foveated image/video quality assessment

A. FWSNR : Foveal Weighted Signal-to-Noise Ratio

The most important HVS attribute is the frequency contrast sensitivity function (CSF). In order to capture the spatially-varying response of the HVS, a local bandlimited contrast has been introduced [2]. In [3], the CSF is $C(f_r) = 2.6(0.0192 + 0.114 f_r) exp[-(0.114 f_r)^{1.1}]$ where f_r is the radial angular frequency (cycles/degree). In [4], the visual sensitivity is dropped in the diagonal directions and the angular frequency f_r is modified by $f_r^{'} = f_r / s(\phi)$ where ϕ is the angle measured from the x axis and $s(\phi)$ is given by $s(\phi) = \frac{1-w}{2} cos(4\phi) + \frac{1+w}{2}$ where the symmetry parameter $w = 0.7$. In [2][5], the CSF was used as a weighting function for noise measurement and the error measurement criterion is the WSNR (weighted SNR) : WSNR $= 10 \log_{10} \frac{\sum_{n=1}^{N} [o(\mathbf{x}_n) * c(\mathbf{x}_n)]^2}{\sum_{n=1}^{N} [(o(\mathbf{x}_n) - r(\mathbf{x}_n)) * c(\mathbf{x}_n)]^2}$ where $*$ denotes linear convolution and $c(\mathbf{x}_n)$ is the CSF in the spatial domain.

The foveal weighting metric f_n^2 is adaptable to other visual quality metrics for localized visual quality assessment. The metric can be simply applied to the WSNR. The quality metric is defined as the FWSNR : FWSNR $= 10 \log_{10} \frac{\sum_{n=1}^{N} [o(\mathbf{x}_n) * c(\mathbf{x}_n)]^2 f_n^2}{\sum_{n=1}^{N} [(o(\mathbf{x}_n) - r(\mathbf{x}_n)) * c(\mathbf{x}_n)]^2 f_n^2}$.

B. Visual Quality Measurement for Additive Noise Images

Fig. 2 depicts images that have been corrupted by (a) Gaussian white noise; (b) by highpass Gaussian noise; (c) by spatially weighted Gaussian white noise, which is densest at the foveation point; and (d) by spatially weighted highpass Gaussian noise which is most sparse at the foveation point and denser towards the periphery. The WSNR and the FWSNR are measured at the visual distance 50 cm. Fig. 2 (b) looks better than Fig. 2 (a) even though the SNR for both images is the same. Thus, the WSNR is well matched with the perceptual visual quality according

to the frequency response of the CSF. However, the weighting is derived from the CSF over the frequency domain, so it cannot adequately quantify spatially-varying distortions as shown in Fig. 2 (c) and (d). Conversely, the FSNR effectively measures spatially-varying additive noise. However, the FSNR cannot adequately quantify distortions occurring in the frequency domain. Thus, the FSNR in (a) and (b) is approximately same despite the different apparent visual quality. The FWSNR overcomes the drawbacks of the FSNR and the WSNR. Using the FWSNR, it is possible to measure localized frequency noise and localized spatial noise.

C. Visual quality measurement for video processing

We have incorporated the FMSE into an optimal rate control algorithm designed using a Lagrange multiplier approach, which yields higher (foveated) visual quality [6]. In the simulations to follow, we used the H.263 video standard. Using a constant quantization parameter, we obtained a standard compressed image in Fig. 3 (a), and a foveated and compressed image in Fig. 3 (c). At equivalent bitrates we applied optimal rate control to the original image and the foveated image, obtaining the reconstructed images in Fig. 3 (b) and (d). It is apparent that the quality measurements made by the PSNR and the WSNR do not effectively quantify the localized visual quality of the foveated and original reconstructed images, presuming that the visual fixation point of the viewer is at the center of foveation. Using optimal bit allocation over curvilinear coordinates, the FPSNR improved by 0.8 dB for the standard compressed image, and by 1.5 dB for the foveated compressed image. Due to foveation filtering, the PSNR was improved by 4.1 dB and the FPSNR by 3.1 dB from image (a) to image (c). When the FWSNR was used, high frequency errors in the hair were reduced by the local bandwidth weighting over the spatial domain, and by the contrast sensitivity weighting over the frequency domain. Thus, the FWSNR of the images is apparently well matched by the subjective visual quality. While the FPSNR effectively measures localized visual quality over the spatial domain, the FWSNR performs simultaneously well over the spatial and the frequency domains. Using optimal rate control based on the FMSE, it is possible to improve foveated visual quality as measured by the FWSNR.

IV. Conclusions

In this paper, we developed new methods for foveated image/video quality assessment. In order to achieve the main goal, we defined a new objective quality criterion (FPSNR/FWSNR) defined

on curvilinear coordinates and based on the foveation response of the human visual system. The novelty of the paper was the introduction of a unique visual quality criterion that utilizes a non-uniform resolution weighting metric. The envision that this new approach to visual quality assessment will find extensive use in the fields of multimedia visual communications, virtual reality, wireless video, web-based applications, virtual 3-D games and so on, as foveated video processing algorithms become more prevalent.

References

[1] B. A. Wandell, *Foundations of Vision.* Sunderland, MA: Sinauer Associates, Inc, 1994.

[2] N. D.-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Processing*, vol. 9, pp. 636–650, April 2000.

[3] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Inform. Theory*, vol. 20, pp. 525–536, July 1974.

[4] J. Sullivan, R. Miller, and G. Pios, "Image halftoning using a visual model in error diffusion," *J. Opt. Soc. Amer.*, vol. 10, pp. 1714–1724, Aug. 1993.

[5] T. D. Kite, B. L. Evans, and A. C. Bovik, "A fast, high-quality inverse halftoning algorithm for error diffused halftones," *IEEE Trans. Image Processing*, vol. 9, pp. 1583–1592, Sep. 2000.

[6] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Trans. Image Processing*, accepted.

(a) $v(\mathbf{x}_n) \in B^{w_f(\mathbf{x})}$

(b) $z(\Phi(\mathbf{x}_n)) \in B^{w_c}$, $w_c = \pi$

Fig. 1. Foveated image in cartesian coordiates (a) and curvilinear coordinates (b)

(a) SNR = 16.0, FSNR = 16.3, WSNR = 16.5, FWSNR = 16.8

(b) SNR = 16.0, FSNR = 16.3, WSNR = 17.0, FWSNR = 17.2

(c) SNR = 16.0, FSNR = 15.2, WSNR = 16.5, FWSNR = 15.7

(d) SNR = 16.0, FSNR = 18.1, WSNR = 17.0, FWSNR = 19.1

Fig. 2. Quality assessment for additive noise images : viewing distance 50 cm

(a) PSNR = 29.9, FPSNR = 29.0, WSNR = 23.1, FWSNR = 22.6

(b) PSNR = 29.0, FPSNR = 29.8, WSNR = 22.3, FWSNR = 23.3

(c) PSNR = 34.0, FPSNR = 32.1, WSNR = 27.1, FWSNR = 25.6

(d) PSNR = 33.2, FPSNR = 33.6, WSNR = 26.2, FWSNR = 26.9

Fig. 3. Quality assessment for the H.263 video coding : I-frame, angular freq. = 8.5 cyc/deg, (a) 35.4 Kbits, (b) 34.6 Kbits, (c) 34.2 Kbits, (d) 35.6 Kbits