Foveated Vision and Video Quality Evaluation

Mario Vranješ, Snježana Rimac-Drlje, Marijan Herceg
Faculty of Electrical Engineering
University of Osijek
Kneza Trpimira 2B, Osijek, HR-31000 Croatia
Phone: +385 91 224 6060 Fax: + 385 31 224 605 E-mail: rimac@etfos.hr

Abstract— **The using of characteristics of the human visual system (HVS) can significantly improve the efficiency of video coding process. Enhanced coding processes take into account the high spatial variability of HVS in sampling, coding and understanding of visual information. The highest visual sensitivity is at the point of fixation and decreases rapidly with distance from that point. Foveated image and video coding systems use the advantage of this phenomenon and therefore achieve increased compression efficiency. The basic idea used in such systems is to remove considerable high-frequency information redundancy from the regions away from the fixation point without significant loss of the reconstructed image or video quality. In this paper we analyze the usage of some foveated techniques in video quality evaluation process. The quality measures obtained by objective quality metrics are compared with the results obtained by subjective tests**.

Keywords— foveated vision, objective video quality, subjective video quality

I. INTRODUCTION

In modern digital communication systems an increasing number of services include digital video as their important part (e.g. video-streaming, videoconference, video surveillance). Video parameters such as resolution, frame rate and bit rate highly depend on transmission channel capacity since quality of video materials has to satisfy expectations of a common user. Because of limited transmission channel capacity, the source video materials have to be compressed. Thus, the enhanced coding efficiency can enable video transmission of higher quality within the existing channel capacities. However, increasing compression rates cause visible compression artifacts highly dependant on video content. Hereafter, transmission errors, common in mobile systems, as well as packet losses, common in IP networks, cause additional video quality degradation. Therefore, an automatic video quality control is needed.

Because of their simplicity, the mean-squared error (MSE) and the peak signal-to-noise ratio (PSNR) are widely used objective video quality metrics. However, these metrics usually cannot give an objective quality measure corresponding well with the quality perceived by a human observer for a wide range of coding and transmission parameters. In recent years there has been an increasing interest in developing a human vision-based objective models for evaluation of image and video quality. A good example of such objective metrics are VQM [1] and SSIM [2].

The HVS is highly space-variant in sampling, processing and understanding of visual information. The reason of this fact is that the photoreceptors (cones and rods) and ganglion cells are non-uniformly distributed in the retina in the human eye [3]. Thus, for example, the spatial resolution of the HVS is highest around the point of fixation (foveation point) and decreases dramatically with increasing eccentricity. The idea used in a foveation image and video processing is to utilize such characteristics and remove considerable high-frequency information redundancy in the peripheral regions. This provides much more efficient representation of image and video files, because the peripheral high-frequency information, that human eye cannot perceive, are truncated.

In this paper we introduce an objective quality metric called FMSE (Foveated MSE) that utilize a non-uniformity of HVS contrast sensitivity, and compare its results with the results of four other objective metrics (MSE, PSNR, VQM and SSIM).

II. FOVEATED VISUAL SENSITIVITY MODEL

Human visual perception is characterized by a variable resolution across the viewing field. The highest resolution occurs at and near the point of fixation and deceases away from these point, as a function of eccentricity, because of the non-uniform distribution of photoreceptors on the retina [6]. The fixation point is projected onto the fovea – the area of densest sampling. Because of that, the overall variable resolution data is called foveated image.

Psychological experiments had been conducted to measure the contrast sensitivity of HVS as a function of retinal eccentricity. In [3], a model that fits the experimental data was given as:

$$CT(f,e) = CT_0 \exp\left(\alpha f \frac{e+e_2}{e_2}\right) \qquad (1)$$

where f is the spatial frequency (cycles/degree), e is the retinal eccentricity (degrees), CT_0 is a constant minimal contrast threshold, α is the spatial frequency decay constant, e_2 is the half-resolution eccentricity, and $CT(f,e)$ is the visible contrast threshold as a function of f and e. The best fitting parameter values given in [3] are $\alpha = 0.106$, $e_2 = 2.3$ and $CT_0 = 1/64$. The contrast sensitivity is than defined as:

$$CS(f,e) = \frac{1}{CT(f,e)} \quad (2)$$

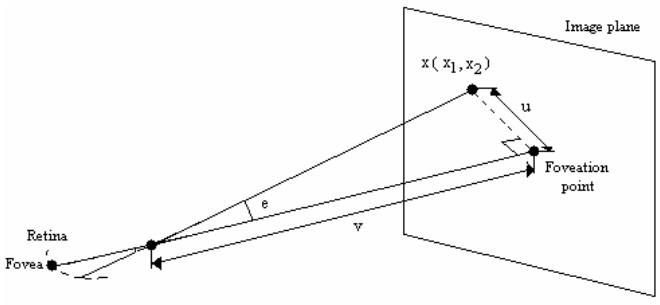
The results presented in [4] show that the contrast sensitivity decreases very fast with increasing eccentricity.

From equation (1) it can be found a critical frequency or cutt-off frequency f_c, as a function of e, by setting CT to 1.0 (the maximum possible contrast):

$$f_c = \frac{e_2 \ln(1/CT_0)}{(e+e_2)\alpha} \text{ (cycles/degree)} \quad (3)$$

Fig 1. shows the typical viewing geometry. It's assumed that the observed image is N pixel wide and the line from the fovea to the fixation point in the image is perpendicular to the image plane. The position of foveation point is $x^f = (x_1^f, x_2^f)^T$ (pixels) and the viewing distance is v (measured in image width) [3].

Figure 1. A typical viewing geometry



The distance u (measured in image width) from point $\boldsymbol{x}$ to $\boldsymbol{x}^f$ is then

$$u = d(\boldsymbol{x}) / N \quad (4)$$

where

$$d(x) = \left\| x - x^f \right\| = \sqrt{(x_1 - x_1^f)^2 + (x_2 - x_2^f)^2} \quad (5)$$

measured in pixels. The eccentricity is given by

$$e(v,x) = \tan^{-1}\left(\frac{u}{v}\right) = \tan^{-1}\left(\frac{d(x)}{Nv}\right) \quad (6)$$

The maximum perceived resolution (in real world images) is limited by the display resolution r:

$$r = \frac{\pi N v}{180} \text{ (pixels/degree)} \quad (7)$$

So the highest frequency that can be represented without aliasing by the display (according to the sampling theorem), or the display Nyquist frequency, is the half of r

$$f_d = \frac{r}{2} = \frac{\pi N v}{360} \text{ (pixels/degree)} \quad (8)$$

Combining (3) and (8) it can be obtained the cutt-off frequency for a given location **x** by

$$f_m(\boldsymbol{x}) = min(f_c, f_d) \quad (9)$$

Finally, the foveation-based error sensitivity for given viewing distance v, frequency f and location $\boldsymbol{x}$ is defined as:

$$S_f(v,f,x) = \begin{cases} \frac{CS(f,e(v,x))}{CS(f,0)} = \exp\left(-\alpha f \frac{e(v,x)}{e_2}\right) & for\ f \le f_m \\ 0 & for\ f > f_m \end{cases} \quad (10)$$

S_f is normalized so that highest value is always 1.0 at eccentricity 0 [6].

III. OBJECTIVE VIDEO QUALITY METRICS

A. PSNR and MSE

PSNR is very popular objective quality metric used both for video and still images. PSNR is defined as

$$PSNR = 10 \log_{10} \frac{(2^M - 1)^2}{MSE} \quad (11)$$

where $2^M - 1$ is the maximum value that pixel can take for M-bit image. MSE is the Mean Squared Error that is defined for a video with frame of size X x Y pixels as

$$MSE = \frac{1}{XYT} \sum_{t=1}^{T} \sum_{y=1}^{Y} \sum_{x=1}^{X} [p(x,y,t) - p'(x,y,t)]^2 \quad (12)$$

Original video frame is presented by p(x,y,t) and distorted one with p'(x,y,t). We have used PSNR for luma component only, whereas color information has not been taken into account.

B. DCT-based VQM metric

VQM is a DCT based video quality metric, developed by F. Xiao, [1], as a simplified version of Watson's Digital Video Quality metric [5]. The VQM is based on simplified human spatial-temporal contrast sensitivity model and it uses a Discrete Cosine Transform in calculation the distortion of a compressed video. More details about the VQM calculation process can be found in [1]. VQM score decreases as the quality of perceived video rises and it is zero for the lossless compressed video.

C. Structural Similarity (SSIM) index

SSIM metric uses structural distortion in video as an estimate of perceived visual distortion, [2]. For the structural distortion measure, SSIM uses means, variances and covariance of original and distorted sequences. It is based on assumption that HVS is highly adapted for extraction the structural information from the viewing field. Also it assumes that the level of perceived impairment is proportional to the perceived structural information loss instead of perceived errors. The more details about SSIM calculation process can be found in [2]. SSIM score increases as the quality of perceived video rises and it is 1 for the lossless compressed video.

D. Foveated MSE (FMSE)

FMSE uses a foveation-based error sensitivity described with (10). For a given display Nyquist frequency f_d described with (8), K different 2D filters are generated to filter the difference between original and distorted frame of the video sequence into K frequency subbands. Central frequencies, f_1, …

f_K, of this filters are uniformly distributed across the area from 0 to f_d, so each filter has a bandwidth of f_d / K.

The assumption is that the foveation point is in the center of the frame. Based on this assumption there is calculated the eccentricity e for each pixel of the frame by using the expression (6) for a given image width N and viewing distance v. By using this eccentricity matrix e, K different error sensitivity matrices, S_{f1}, … $S_{fK,}$ are calculated for K different central frequencies f_1, … f_K.

In the first step the absolute difference between original and distorted frame is filtered with K different filters and K different filtered versions, $F_1,\ldots,F_K$, of this difference are obtained. Each element of $F_1,\ldots,F_K$, is then multiplied with the appropriate element in the corresponding error sensitivity matrix S_{f1}, … S_{fK}. After this process there are K new matrices $FS_1,\ldots FS_K$. In these matrices each element is squared and for each matrix the sum of these squared elements is divided by overall number of elements (M x N). So, K different numbers, $NUM_1,\ldots, NUM_K$ are obtained. Then the foveated MSE for each frame is calculated as:

$$FMSE = NUM_1 + \ldots + NUM_K \quad (13)$$

The FMSE score decreases as the quality of perceived video rises and it is zero for the lossless compressed video.

IV. EXPERIMANTAL RESULTS AND ANALYSIS

In our experiments we have made several objective and subjective tests for sequences coded by two different codecs. The first codec is the JVT JM version 10.2 as the implementation of the H.264/AVC standard. We have used Main profile and L2 level with CABAC coding and rate-distortion optimization. The second codec is the open source codec XviD version 1.1.0, based on the MPEG-4 Visual standard. The codec has worked in Advance simple profile and L4 level with an adaptive quantization, quarter pixel motion vector accuracy, global motion compensation and bidirectional prediction.

Four progressive video sequences have been coded in our experiment. We have used the *Container*, *Foreman*, *Hall* and *News* sequences all in CIF (352x288) resolution with 25 fps (frames per second) with total of 300 frames. These sequences have been coded at eight different coding rates from 64 to 768 kb/s. The choosing of sequences is based on their very distinct content, the short description of which is given in TABLE 1.

TABLE I. CONTENT DESCRIPTION OF SEQUENCES

Name	**Characteristics**
Container	Slow moving container ship, stationary camera
Foreman	Talking head, with pan to construction site, geometric shapes, shaking camera
Hall	An example of video supervision, stationary camera, two moving persons
News	Male and female speaker in newsroom. movement in background

We have made the subjective quality evaluation for 15 non-experienced observers by using MSU Perceptual Video Quality Tool, [7], and the Subjective Assessment Method for Video Quality evaluation (SAMVIQ). The SAMVIQ is an Absolute Category Rating method referring to the ITU-T P-910 Recommendation for subjective video quality assessment of multimedia application, [8]. More detailed information about subjective measurement process can be found in [9]. Results of subjective measurements are expressed in Mean Opinion Score (MOS) given as an average for all observers.

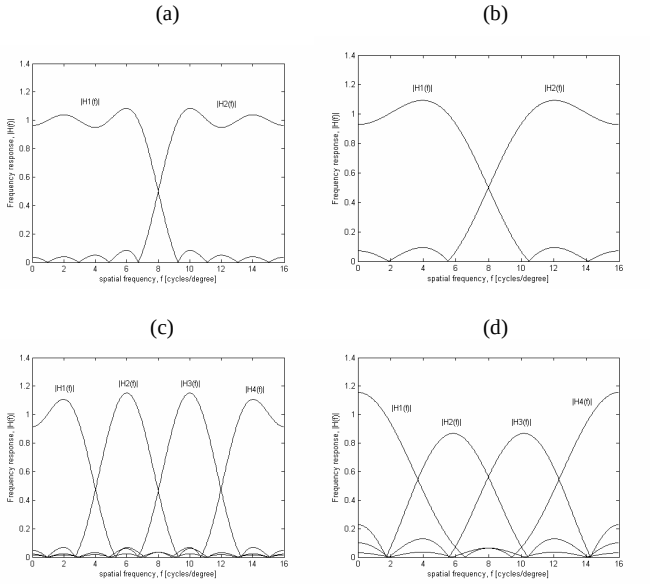
Five objective metrics are used in our experiments: MSE, PSNR, VQM, SSIM and FMSE. MSE, PSNR, VQM and SSIM are obtained by using MSU Video Measurement Tool, [7].

FMSE was calculated for a viewing distance v = 5. Thus the display Nyquist resolution, according to (8) is approximately 16 cycles/degree. Three different groups of filters have been generated:

- in the first group K is equal to 2 and central frequencies of the filters are f_1 = 4 and f_2 = 12 cycles/degree (Fig. 2. (a))
- in the second group K is equal to 4 and central frequencies are f_1 = 2, f_2 = 6, f_3 = 10, f_4 = 14 cycles/degree (Fig. 2. (c));
- in the third group K is equal to 8 and central frequencies are f_1 = 1, f_2 = 3, f_3 = 5, f_4 = 7, f_5 = 9, f_6 = 11, f_7 = 13, f_8 = 15 cycles/degree.

These are FIR filters, with order equal to 14 and Kaiser window with Beta parameter equal to 0.5. Also, the same filter groups were generated but with order equal to 8 (Fig. 2. (b) and (d)).

Figure 2. Frequency response of the filters for a given number of filters and given order of filters (a) K = 2, order = 14 (b) K = 2, order = 8 (c) K = 4, order = 14 (d) K= 4, order = 8



As a measure of metric's accuracy we calculated Pearson linear correlation coefficient, most commonly used for video quality metric evaluation, [10]. For a set of Z data pairs (x_i, y_i) it is defined as follows:

$$r_p = \frac{\sum_{i=1}^{Z}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{Z}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{Z}(y_i - \bar{y})^2}} \quad (14)$$

MOS results are used as y_i and scores from given metric are used as x_i.

In TABLE 2. Pearson correlation coefficients are given as a measure for correlation between MOS and metrics: MSE; PSNR, VQM, SSIM and FMSE. Results are given for sequences coded by each codec alone, as well as for all sequences together.

TABLE II. PEARSON CORRELATION COEFFICIENTS BETWEEN MOS AND MSE, PSNR, VQM, SSIM, FMSE

Metric	r_p for sequences coded by XviD	r_p for sequences coded by JVT JM	r_p for all sequences
MSE	-0.834	-0.937	-0.891
PSNR	0.777	0.885	0.821
VQM	-0.904	-0.968	-0.895
SSIM	0.820	0.935	0.873
FMSE (K=2, order = 14)	-0.857	-0.940	-0.904
FMSE (K=4, order = 14)	-0.813	-0.935	-0.880
FMSE (K=8, order = 14)	-0.803	-0.941	-0.876
FMSE (K=2, order = 8)	-0.859	-0.940	-0.904
FMSE (K=4, order = 8)	-0.814	-0.940	-0.882
FMSE (K=8, order = 8)	-0.800	-0.939	-0.874

The best correlation is obtained by FMSE metric with K=2, order=14 and order=8 (r_p = -0.904). For all sequences Pearson coefficient r_p is -0.895 for VQM, -0.891 for MSE, 0.873 for SSIM and 0.821 for PSNR. The FMSE metric outperforms both the VQM and the SSIM, despite the fact that FMSE includes only one significant characteristic of HVS. It could be concluded that FMSE is a better choice for a objective video quality evaluation, at least for the CIF video sequences. Particularly, for a lower number of filters used and for a lower order of the filters, calculation complexity of the FMSE is significantly reduced in comparison with VQM and SSIM. In comparison with MSE and PSNR it could be seen that FMSE provides a results that correlate better with subjective results and the reason is the usage of foveation-based contrast error sensitivity.

These preliminary results for different filter groups show that the best correlation is obtained by K=2, when only two filters were used. If the number of the used filters increases, the Pearson coefficient decreases for both filter order. It is interesting to note that the filter order didn't play a significant role in a overall score and it can be concluded that a FMSE is a robust method in terms of filter design process.

V. CONCLUSION

An effort to find a good objective video quality metric gives rise to development of complex and quite accurate metrics. A drawback of this metrics is a calculation complexity that is sometimes not acceptable. Also, these metrics don't use some important characteristics of HVS. The FMSE uses the foveation-based contrast error sensitivity and shows a high potential of being a simple metric with good prediction capacity. It uses a group of filters in calculation process. Thus , the filter design process and the behavior of FMSE when the foveation point is not in the center of the frame will be the subject of our further research.

REFERENCES

[1] F.Xiao,"DCT-based Video Quality Evaluation", http://www.compression.ru/video/quality_measure/vqm.pdf

[2] Z.Wang, L. Lu, A.C. Bovik, "Video Quality Assessment Based on Structural Distortion Measurement", Signal Processing: Image Comm. Vol.19, 2004, pp 121-132

[3] H.R.Wu, K.R. Rao (ed.), "Digital Video Image Quality and Perceptual Coding", Taylor &Francis Group, LLC, 2006.

[4] S. Lee, M.S. Pattichis, A.C.Bovik, "Foveated Video Quality Assesment", IEEE Transactions on Multimedia, vol.4, no. 1, march 2002, pp. 129-132

[5] A.B. Watson, J.Hu, J.F. McGowan III, "DVQ: A digital video quality metric based on human vision" , J. Electronic Imaging, vol.10, no. 1, 2001, pp. 20-29

[6] Z.Wang, A.C. Bovik, L. Lu, J. Kouloheris, "Foveated Wavelet Image Quality Index", Proceedings of SPIE Vol. 4472, 2001., pp. 42-52

[7] MSU Graphics&Media Lab, Video Group, MSU codecs, www.compression.ru/video/

[8] Rec. ITU-R BT 500-11, "Methodology for the subjective assessment of the quality of television pictures", 2002

[9] M. Vranješ, S.Rimac-Drlje, K.Grgić, "Locally Averaged PSNR as a Simple Objective Video Quality Metric", Proceedings ELMAR-2008, Vol. 1 of 2, 2008, pp. 17-20

[10] S.Winkler, "Digital video quality: vision models and metrics", Wiley, Chichester, 2005