Foveated Wavelet Image Quality Index*

Zhou Wang[a], Alan C. Bovik[a], and Ligang Lu[b]

[a]Laboratory for Image and Video Engineering (LIVE), Dept. of Electrical and Computer Engineering
The University of Texas at Austin, Austin, TX 78712-1084, USA
[b]Video and Image Systems, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
E-mail: zwang@ece.utexas.edu, bovik@ece.utexas.edu, lul@us.ibm.com

ABSTRACT

The human visual system (HVS) is highly non-uniform in sampling, coding, processing and understanding. The spatial resolution of the HVS is highest around the point of fixation (foveation point) and decreases rapidly with increasing eccentricity. Currently, most image quality measurement methods are designed for uniform resolution images. These methods do not correlate well with the perceived foveated image quality. Wavelet analysis delivers a convenient way to simultaneously examine localized spatial as well as frequency information. We developed a new image quality metric called foveated wavelet image quality index (FWQI) in the wavelet transform domain. FWQI considers multiple factors of the HVS, including the space variance of the contrast sensitivity function, the spatial variance of the local visual cut-off frequency, the variance of human visual sensitivity in different wavelet subbands, and the influence of the viewing distance on the display resolution and the HVS features. FWQI can be employed for foveated region of interest (ROI) image coding and quality enhancement. We show its effectiveness by using it as a guide for optimal bit assignment of an embedded foveated image coding system. The coding system demonstrates very good coding performance and scalability in terms of foveated objective as well as subjective quality measurement.

Keywords: image quality assessment, human visual system (HVS), foveation, wavelet, image coding

1. INTRODUCTION

The photoreceptors (cones and rods) and ganglion cells are non-uniformly distributed in the retina in the human eye [1]. The density of cone receptors and ganglion cells plays important role in determining the ability of our eyes to resolve what we see. Spatially, the visual resolution is highest around the point of fixation (foveation point) and decreases rapidly as a function of eccentricity. Consequently, the human visual system (HVS) is highly spatial-variant in sampling, coding, processing and understanding visual information. The motivation behind foveation image processing is that there exists considerable high frequency information redundancy in the peripheral regions, thus a much more efficient representation of images can be obtained by removing or reducing such information redundancy, provided the foveation point(s) and the viewing distance can be discovered. There has been growing recent interest in research work on foveated image processing [2-5], including foveation filtering and foveated image and video compression.

Currently, most image quality measurement methods are designed for uniform resolution images. However, little has been done in the assessment of non-uniform resolution images. For example, peak signal-to-noise ratio is still used for region of interest (ROI) image coding and postprocessing [6, 7]. Quality assessment method is very important for foveated image coding, because image coding is essentially an optimization procedure that attempts to maximize image quality with a limited number of bits, where the quality metric servers as a guide for bit assignment. Quality metrics are

* This research was partially supported by IBM Corporation, Texas Instrument, Inc., and Texas Advanced Technology Program.

also very useful for the design of quality enhancement algorithms for the preprocessing and postprocessing of images. Wavelet analysis delivers a convenient way to simultaneously examine localized spatial as well as frequency information. In this paper, we present a foveation-based HVS image quality metric, namely foveated wavelet image quality index (FWQI), in the discrete wavelet transform (DWT) domain.

2. FOVEATED VISUAL SENSITIVITY MODEL

Let us first examine the anatomy of the early vision system. The light first passes through the optics of the eye and is then sampled by the photoreceptors on the retina. There are two kinds of photoreceptors – cones and rods. The cone receptors are responsible for daylight vision. Their distribution is highly non-uniform on the retina. The density of the cone cells is highest at the fovea and drops very fast with increasing eccentricity. The photoreceptors deliver data to the bipolar cells, which in turn supply information to the ganglion cells. The distribution of ganglion cells is also highly non-uniform. The density of the ganglion cells drops even faster than the density of the cone receptors. These density distributions play important roles in determining the resolution ability of the human eye. Psychological experiments had been conducted to measure the contrast sensitivity as a function of retinal eccentricity [2, 8-9]. In [2], a model that fits the experimental data was given as

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

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 [2] are $\alpha = 0.106$, $e_2 = 2.3$, and $CT_0 = 1/64$. It was also reported in [2] that the same values of a and e_2 provide a good fit to the data in [8] with $CT_0 = 1/75$, and an adequate fit to the data in [9] with $CT_0 = 1/76$, respectively. We use the parameter selections as in [2]. The contrast sensitivity is defined as:

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

For a given e, equation (1) can be used to find its critical frequency or so called cut-off frequency f_c by setting CT to 1.0 (the maximum possible contrast) and solving for e

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

Assume that the observed image is N pixels wide and the line from the fovea to the point of fixation in the image is perpendicular to the image plane. Also assume that the position of the foveation point $\boldsymbol{x}^f = (x_1^f, x_2^f)^T$ (pixels) and the viewing distance v (measured in image width) from the eye to the image plane are known. The distance u (measured in image width) from point $\boldsymbol{x} = (x_1, x_2)^T$ to $\boldsymbol{x}_f$ is then $u = d(\boldsymbol{x})/N$, where $d(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}^f\|_2 = \sqrt{(x_1 - x_1^f)^2 + (x_2 - x_2^f)^2}$ (measured in pixels). The eccentricity is given by

$$e(v, \boldsymbol{x}) = \tan^{-1}\left(\frac{u}{v}\right) = \tan^{-1}\left(\frac{d(\boldsymbol{x})}{Nv}\right). \tag{4}$$

Fig. 1 shows the normalized contrast sensitivity as a function of pixel position for N = 512 and v = 1, 3, 6 and 10, respectively. The cut-off frequency as a function of pixel position is also given. The contrast sensitivity is normalized so that the highest value is always 1.0 at 0 eccentricity. It can be observed that the cut-off frequency drops quickly with increasing eccentricity and the contrast sensitivity decreases even faster. In real world digital images, the maximum perceived resolution is also limited by the display resolution r:

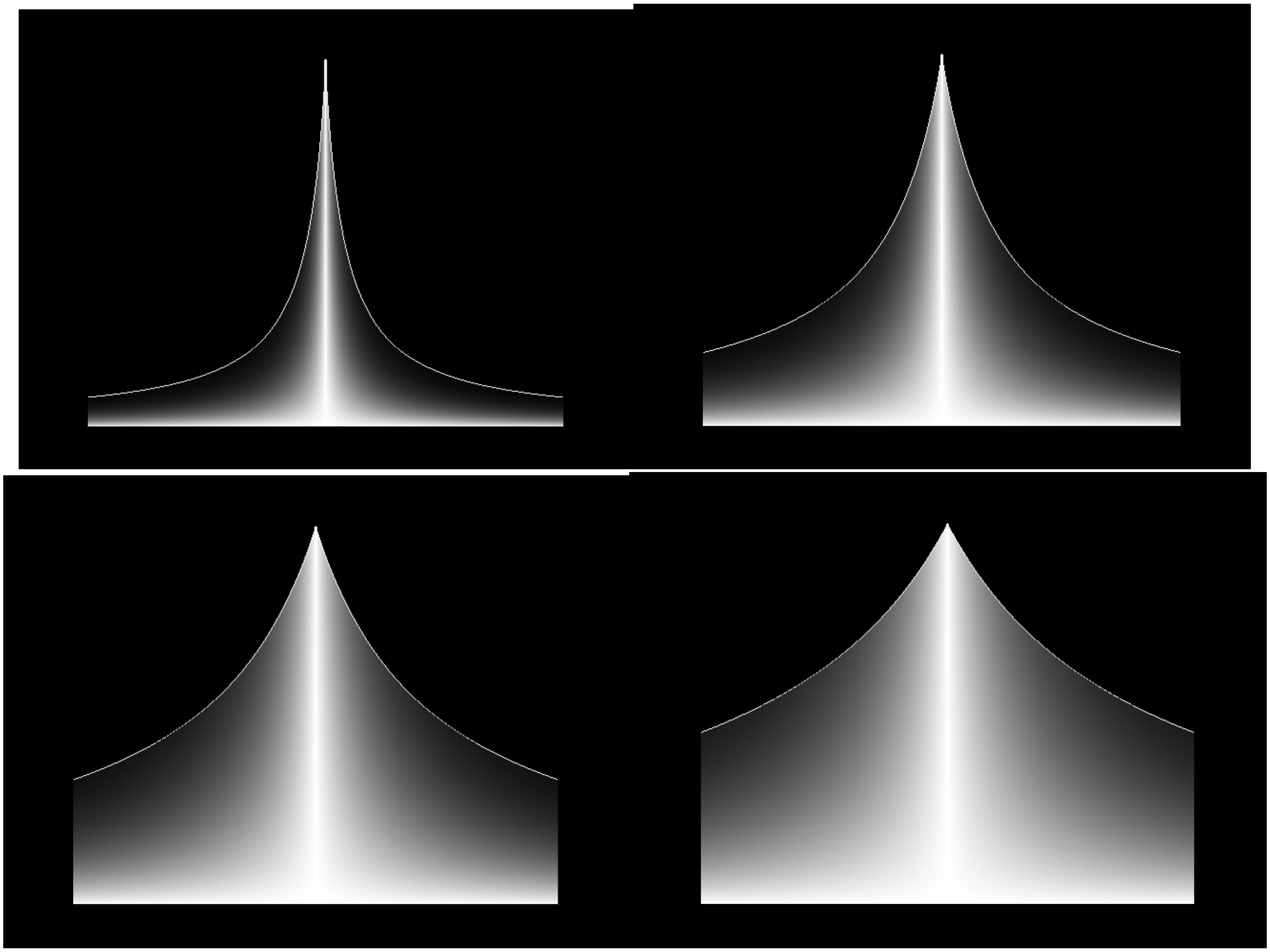


Fig. 1 Normalized contrast sensitivity (Brightness indicates the strength of contrast sensitivity). The top-left, top-right, bottom-left and bottom-right figures are for N = 512 and viewing distance v = 1, 3, 6 and 10 times of the image width, respectively. The white curves show the cutoff frequency.

$$r = \frac{\pi N v}{180}\sec^2\left(\frac{\pi e}{180}\right) = \frac{\pi N v}{180}\cdot\frac{N^2 v^2}{d^2(\boldsymbol{x}) + N^2 v^2} \approx \frac{\pi N v}{180} \text{ (pixels/degree)} \tag{5}$$

This approximation is equivalent to that given in [10]. According to the sampling theorem, the highest frequency that can be represented without aliasing by the display, or the display Nyquist frequency, is half of r:

$$f_d = \frac{r}{2} \approx \frac{\pi N v}{360} \text{(cycles/degree).} \tag{6}$$

Combining (3) and (6), we obtain the cutoff frequency for a given location $\boldsymbol{x}$ by:

$$f_m(\boldsymbol{x}, v) = \min(f_c(e(\boldsymbol{x}, v)),\ f_d(v)) = \min\left(\frac{39.23}{1 + 0.435\tan^{-1}\left(d(\boldsymbol{x})/Nv\right)}, \frac{\pi N v}{360}\right). \tag{7}$$

At a small viewing distance such as v = 1, the display Nyquist frequency is so small that the cutoff frequency stays almost unchanged for a large range of eccentricities. However, strong “foveation” is still obtained because the contrast sensitivity is very sensitive to eccentricity, as shown in Fig. 1. Finally, we define the foveation-based error sensitivity for given viewing distance v, frequency f and location $\boldsymbol{x}$ as:

$$S_f(v, f, \boldsymbol{x}) = \begin{cases} \dfrac{CS(f, e(v, \boldsymbol{x}))}{CS(f, 0)} = \exp(-0.0461 f \cdot e(v, \boldsymbol{x})) & for \;\; f \le f_m(\boldsymbol{x}) \\ 0 & for \;\; f > f_m(\boldsymbol{x}) \end{cases}. \tag{8}$$

S_f is normalized so that the highest value is always 1.0 at 0 eccentricity.

3. FOVEATED IMAGE QUALITY METRIC

The DWT has proved to be a powerful tool for image processing and coding. In the 1-D DWT, the input discrete signal $\boldsymbol{s}$ is convolved with highpass and lowpass analysis filters and downsampled by two, resulting in transformed signals $\boldsymbol{s}_H$ and $\boldsymbol{s}_L$. The signal $\boldsymbol{s}_L$ can be further decomposed and the process may be repeated multiple times. The number of repetitions defines the wavelet decomposition level λ. For image processing, the horizontal and vertical wavelet decompositions are applied alternatively, yielding LL, HL, LH and HH subbands. The LL subband may be further decomposed and the process repeated multiple times. A typical DWT decomposition structure is given by Fig. 2. Let (λ, θ) represent the subband of level λ and orientation θ, where θ is an index representing the LL, LH, HH or HL subband. The wavelet coefficients at different subbands supply information of variable perceptual importance. In [10], psychovisual measurement results were given for the visual sensitivity in wavelet decompositions. A model that fits the experimental data is [10]

$$\log Y = \log a + k(\log f - \log g_\theta f_0)^2, \tag{9}$$

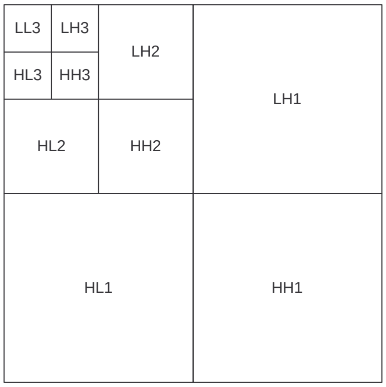


Fig. 2 DWT decomposition structure.

where Y is the visually detectable noise threshold and $f = r2^{-\lambda}$ [10] is the spatial frequency. For gray scale models, a is [10] 0.495, k is 0.466, f_0 is 0.401, and g_θ is 1.501, 1, and 0.534 for the LL, LH/HL, and HH subbands, respectively. The error sensitivity in subband (λ, θ) is given by:

$$S_W(\lambda, \theta) = \frac{A_{\lambda,\theta}}{Y_{\lambda,\theta}} = \frac{A_{\lambda,\theta}}{A_{\lambda,\theta} a 10^{k(\log(2^{\lambda} f_0 g_\theta / r)^2}}, \tag{10}$$

where $A_{\lambda,\theta}$ is the basis function amplitude given in [10]. Let $\boldsymbol{B}_{\lambda,\theta}$ denote the set of wavelet coefficient positions residing in subband (λ, θ). For each subband, we calculate the corresponding foveation point $\boldsymbol{x}^f_{\lambda,\theta}$ in it:

$$\text{LL: } \boldsymbol{x}^f_{\lambda,\theta} = \left(\frac{x^f_1}{2^\lambda}, \frac{x^f_1}{2^\lambda} \right)^T; \qquad \text{LH: } \boldsymbol{x}^f_{\lambda,\theta} = \left(\frac{x^f_1 + N}{2^\lambda}, \frac{x^f_1}{2^\lambda} \right)^T;$$

$$\text{HL: } \boldsymbol{x}^f_{\lambda,\theta} = \left(\frac{x_1^f}{2^\lambda}, \frac{x_1^f + N}{2^\lambda}\right)^T; \qquad \text{HH: } \boldsymbol{x}^f_{\lambda,\theta} = \left(\frac{x_1^f + N}{2^\lambda}, \frac{x_1^f + N}{2^\lambda}\right)^T. \tag{11}$$

Given a wavelet coefficient at $\boldsymbol{x} \in \boldsymbol{B}_{\lambda,\theta}$, its equivalent distance from the foveation point in the spatial domain is given by $d_{\lambda,\theta}(\boldsymbol{x}) = 2^\lambda \left\| \boldsymbol{x} - \boldsymbol{x}^f_{\lambda,\theta} \right\|_2$. With this distance, we have

$$S_f(v, f, \boldsymbol{x}) = S_f(v, r2^{-\lambda}, d_{\lambda,\theta}(\boldsymbol{x})) \ \ for \ \ \boldsymbol{x} \in \boldsymbol{B}_{\lambda,\theta}. \tag{12}$$

A foveation-based error sensitivity model in the DWT domain is obtained by combining (10) and (12):

$$S(v, \boldsymbol{x}) = \left[S_w(\lambda, \theta)\right]^{\beta_1} \cdot \left[S_f(v, r2^{-\lambda}, d_{\lambda,\theta}(\boldsymbol{x}))\right]^{\beta_2} \ \ for \ \ \boldsymbol{x} \in \boldsymbol{B}_{\lambda,\theta}, \tag{13}$$

where β_1 and β_2 are parameters used to control the magnitudes of S_w and S_f, respectively. We use $\beta_1 = 1$ and $\beta_2 = 2.5$. Fig. 3 shows $S(v, \boldsymbol{x})$ for v = 1, 3, 6 and 10, respectively. We define a foveated wavelet image distortion (FWD) metric as:

$$FWD = \left(\frac{1}{M} \sum_{n=1}^{M} \left[S(v, \boldsymbol{x_n}) \cdot \left| c(\boldsymbol{x_n}) - c'(\boldsymbol{x_n}) \right|\right]^Q\right)^{1/Q}, \tag{14}$$

where M is the number of the wavelet coefficients, and $c(\boldsymbol{x_n})$ and $c'(\boldsymbol{x_n})$ are the n-th wavelet coefficients of the original and the compressed images at location $\boldsymbol{x_n}$ in the DWT domain, respectively. Q is set to 2 in our measurement system. We define a foveated wavelet image quality index (FWQI) as:

$$FWQI = \exp(-FWD). \tag{15}$$

The value of FWQI is between 0 and 1, with the maximum value 1 at FWD = 0. Because $S(v, \boldsymbol{x_n})$ varies with the viewing distance v, both FWD and FWQI of a test image are functions of v, instead of single values.

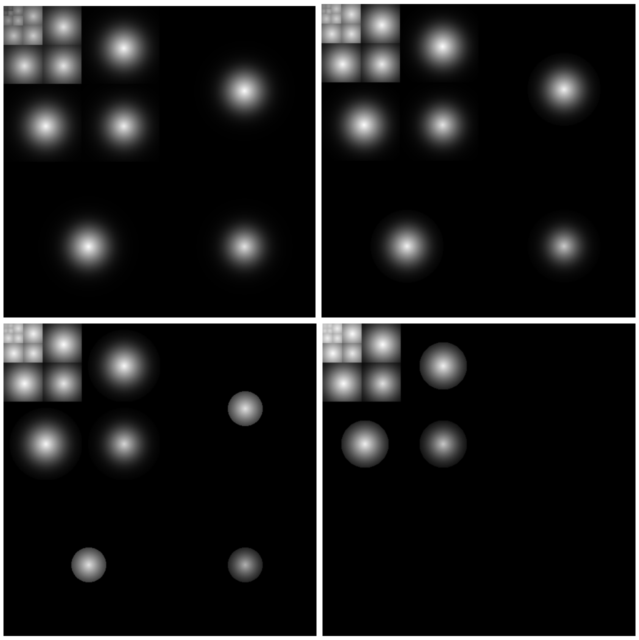


Fig. 3 Foveation-based error sensitivity mask in the DWT domain. The top-left, top-right, bottom-left, and bottom-right figures are for viewing distance v = 1, 3, 6 and 10 times of the image width, respectively. (Brightness logarithmically enhanced for display purpose)

Although there is only one foveation point at one time for one human observer, it is necessary to allow multiple foveation points in practice to provide more flexibility and robustness. This is because 1) the usual pattern of human fixation is that the fixation point moves slightly around a small area of the center point of interest, 2) there may be multiple human observers watching the image at the same time, and 3) there may exist multiple points and/or regions in the image that have high probability to attract a human observer's attention. Our system can easily adapt to multiple foveation points by changing the error sensitivity mask $S(v, \boldsymbol{x})$. Suppose that there are P foveation points $\boldsymbol{x}_1^f$, $\boldsymbol{x}_2^f$, … , $\boldsymbol{x}_P^f$ in the image (in digitally sampled images, the foveation regions can also be regarded as collections of foveation points). For each of the points, we can calculate the error sensitivity mask as in the above sections and have $S_i(v, \boldsymbol{x})$ for $i = 1, 2, \ldots, P$. The overall error sensitivity should be given by the maximum of them:

$$S(v, \boldsymbol{x}) = \max_{i=1\cdots P}(S_i(v, \boldsymbol{x})) \cdot \tag{16}$$

In practice, it is not necessary to compute each $S_i(v, \boldsymbol{x})$. Since the error sensitivity is monotonically decreasing with increasing distance from the foveation point, given a point $\boldsymbol{x}$, the foveation point that is closest to it must generate the maximum $S_i(v, \boldsymbol{x})$, so what we need to do is let

$$S(v, \boldsymbol{x}) = S_j(v, \boldsymbol{x}), \qquad for \ \ j \in \arg\min_{i\in\{1,2,\cdots,P\}}\left\{\left\|\boldsymbol{x} - \boldsymbol{x}_i^f\right\|_2\right\}. \tag{17}$$

By doing this, a large amount of computation is saved.

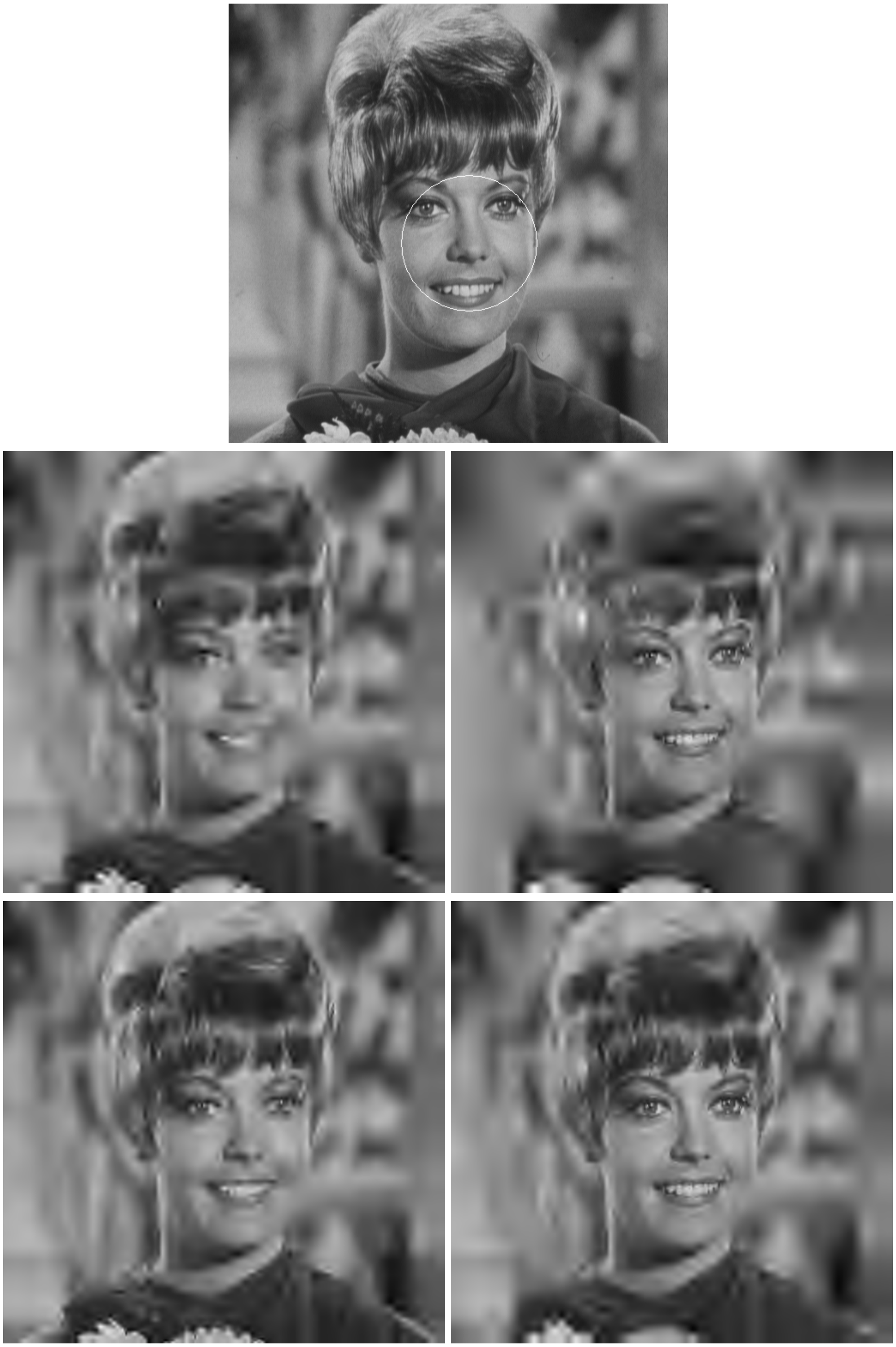
4. IMAGE CODING USING THE FOVEATED QUALITY METRIC

SPIHT [11] is a very efficient progressive wavelet image coding algorithm. We designed a modified SPIHT algorithm and tuned it using the above FWQI model to optimize the foveated visual quality at any given bit rate. We call the new coding algorithm the embedded wavelet image coding (EFIC) algorithm [5]. The encoded bitstream can be truncated at arbitrary place to create reconstructed images with different quality and depth of foveation. Fig. 4 shows the 512×512 "Zelda" image encoded with both SPIHT and EFIC algorithms. Fig. 5 gives the FWQI comparisons of the EFIC and SPIHT compressed "Zelda" images at 0.015265, 0.0625 and 0.25bpp, respectively. The FWQI for each image is given as a function of the viewing distance, instead of just one fixed value. Significant quality gain is achieved throughout the whole range of the viewing distances. This is consistent with the foveated subjective quality of Fig. 4. Fig. 6 shows how the FWQI result changes with the encoding bit rate. In Fig. 7, we compare the 288×352 "News" image compression results at the same bit rate of 0.25bpp but with different ROI region selections. It turns out that uniform resolution SPIHT coding cannot provide an acceptable image, but if the ROIs are known to us, visually satisfactory quality image is still achievable with the EFIC algorithm.

The EFIC decoding can also be viewed as a foveation filtering process with decreasing foveation depth. Note that, in typical natural images, the energy is concentrated in the low frequency bands. As a result, in the peripheral regions, the low frequency wavelet coefficients have greater opportunity to be reached before the high frequency ones. In the region of fixation, both low and high frequency coefficients have good chances to be reached early because of their larger importance weights. If the bit rate is limited, then decoding corresponds to applying all-pass filtering to the region of fixation and low-pass filtering to the peripheral regions. This is the basic idea of foveation filtering. With an increase of the bit rate, more bits are received for the high frequency coefficients of peripheral regions, thus the decoded image becomes less foveated. The EFIC coding results in Fig. 4 demonstrate this very well.

5. CONCLUSION AND DISCUSSION

We described our foveated wavelet image quality measurement approach, which considers multiple factors of the HVS, including the space variance of the contrast sensitivity function, the spatial variance of the local visual cut-off frequency, the variance of human visual sensitivity in different wavelet subbands, and the influence of the viewing distance on the display resolution and the HVS features. We show its effectiveness by using it as a guide for optimal bit assignment of an embedded foveated image coding system.



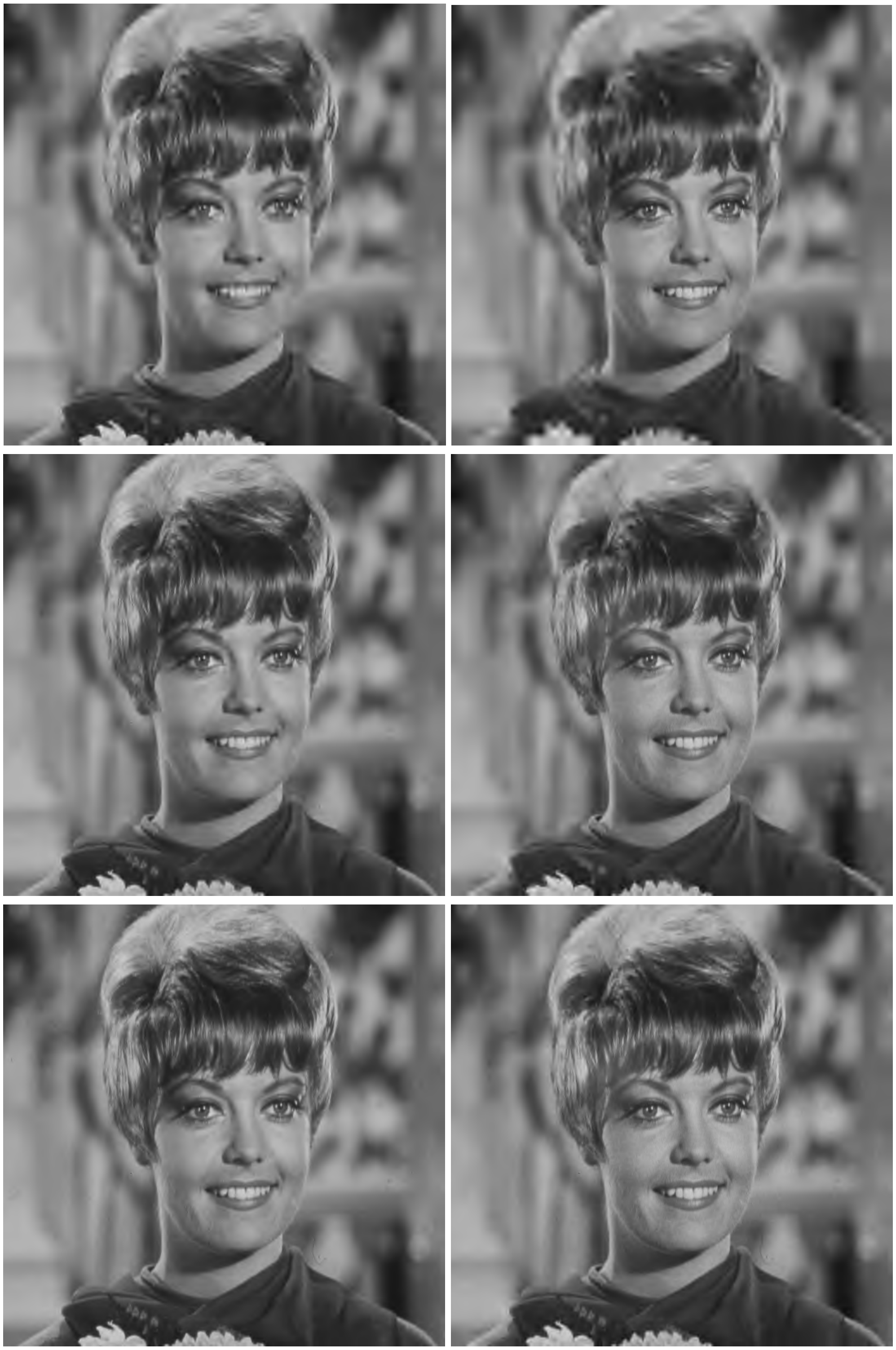


Fig.4 “Zelda” image compression result comparison. Top: Original image with the foveated ROI indicated; The left images that followed: SPIHT coded images; The right images that followed: EFIC coded images. The bit rates from top to bottom are 0.015625bpp (CR=512:1), 0.03125bpp (CR=256:1), 0.0625bpp (CR=128:1), 0.125bpp (CR=64:1), and 0.25bpp (CR=32:1), respectively.

When we introduce our foveation image coding and processing work to people, the most frequently asked question is: "How do you know the foveation points?" Generally, there are two methods to determine the fixation point(s) and region(s). The first is a completely automatic method. There has been a lot of research work in the visual psychology community towards understanding high level and low level processes in deciding human fixation points [12, 13]. High level processes involves a cognitive understanding of the image. For example, once a human face is recognized in an image, the face area is very likely to become a heavily fixated region. Low level processes determine the points of interest using simple local features of the image [13]. The second method to determine foveation point(s) is the interactive method. In some applications, an eye tracker is available, which can track the fixation point and send it to the foveated imaging system in real time. In some other application environments, the eye tracker is not available or inconvenient. A more practical way is to ask the users to indicate fixation points using a mouse. Another practical possibility is to ask the users to indicate the object of interest, and an automatic algorithm is employed used to track the user-selected object as the foveated region in the image sequence that follows.

It is worth noting that in a foveated system, no object segmentation is needed. As shown in Fig. 4 and Fig. 7, it is not necessary for a foveated system to extract the boundary of an object precisely. A rough foveated region is enough for the foveated system to work properly. Note that manually picking foveation points is much easier than manually defining ROIs in the image. Also note that automatically locating foveation regions is much easier than automatically segmenting objects from the image. In this sense, a foveated image coding and communication system is more implementable, flexible, robust and thus practical than the segmentation-based ROI coding systems.

REFERENCES

[1] L. R. Cormack, "Computational models of early human vision," in *Handbook of image and video processing*, Al Bovik, Ed., Academic Press, May 2000.

[2] W. S. Geisler, and J. S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication," *Proceedings of SPIE*, vol. 3299, 1998.

[3] E.-C. Chang, "Foveation techniques and scheduling issues in thinwire visualization," *Ph.D. Dissertation*, New York University, May 1998.

[4] S. Lee, "Foveated video compression and visual communication over wireless and wireline networks," *Ph.D. Dissertation*, The University of Texas at Austin, May 2000.

[5] Z. Wang and A. C. Bovik, "Embedded foveation image coding," accepted by *IEEE Trans. Image Processing*, 2001.

[6] E. Atsumi, and N. Farvardin, "Lossy/lossless region-of-interest image coding based on set partitioning in hierarchical trees," *IEEE International Conference on Image Processing*, vol. 1, pp. 87-91, 1998.

[7] J. Jung, S. Joung, Y. Jang, and J. Paik, "Enhancement of region-of-interest coded images by using adaptive regularization," *IEEE International Conference on Consumer Electronics*, pp. 62-63, 2000.

[8] T. L. Arnow, and W. S. Geisler, "Visual detection following retinal damage: predictions of an inhomogeneous retino-cortical model," *Proceedings of SPIE: Human vision & Electro. Imaging*, vol. 2674, pp. 119-130, 1996.

[9] M. S. Banks, A. B. Sekuler, and S. J. Anderson, "Peripheral spatial vision: limits imposed by optics, photoreceptors, and receptor pooling," *Journal of the Optical Society of America*, vol. 8, pp. 1775-1787, 1991.

[10] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Processing*, vol. 6, no. 8, pp. 1164-1175, Aug. 1997.

[11] A. Said, and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 6, no. 3, pp. 243-250, June 1996.

[12] A. L. Yarbus, *Eye Movements and Vision*, Plenum press, New York, 1967.

[13] C. M. Privitera and L. W. Stark, "Algorithm for defining visual regions-of-interest: comparison with eye fixations," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 970-982, Sep. 2000.

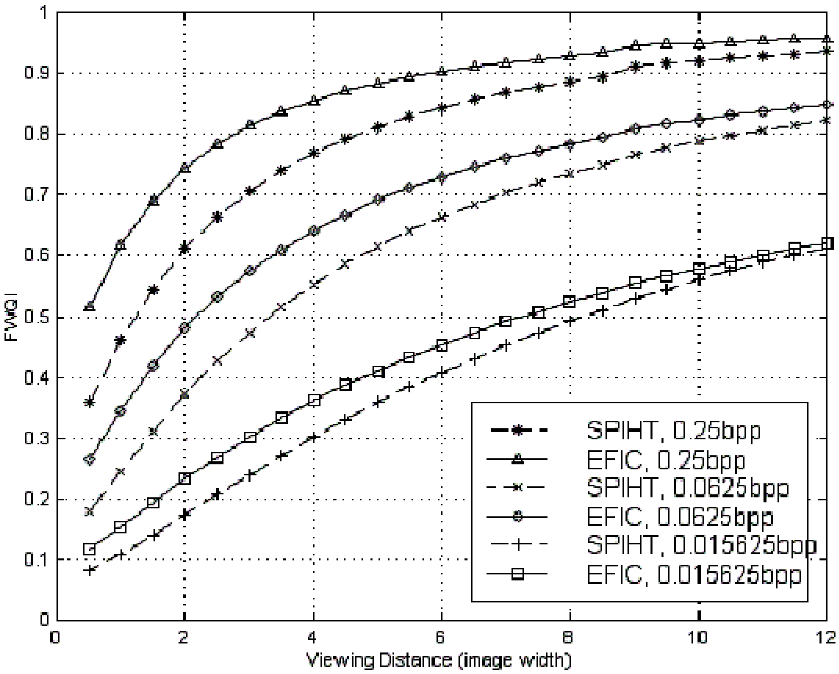


Fig. 5 FWQI comparison of EFIC and SPIHT compressed "Zelda" image at 0.15625bpp, 0.0625bpp and 0.25bpp.

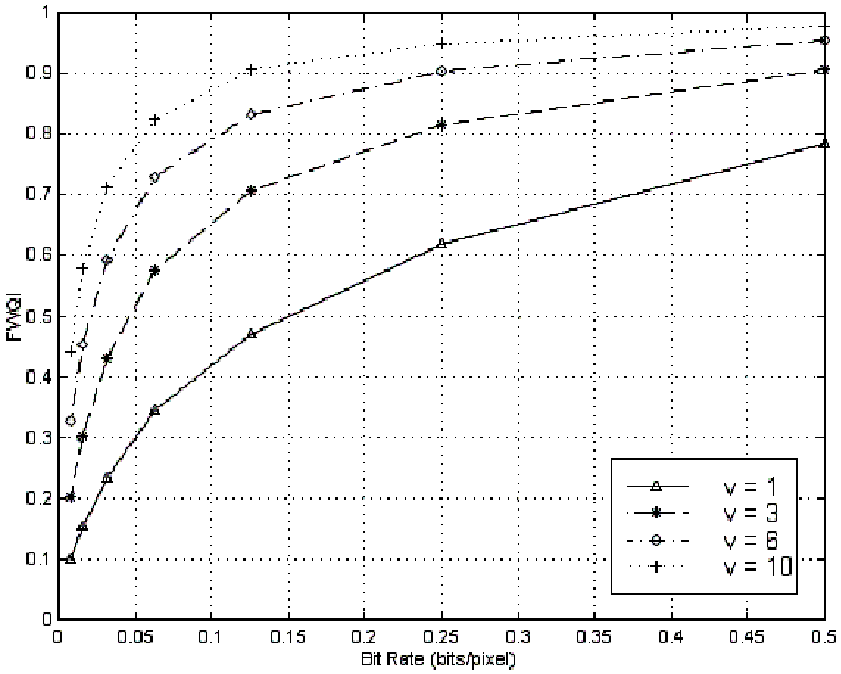


Fig. 6 FWQI results of EFIC compressed "Zelda" image at different bit rates.

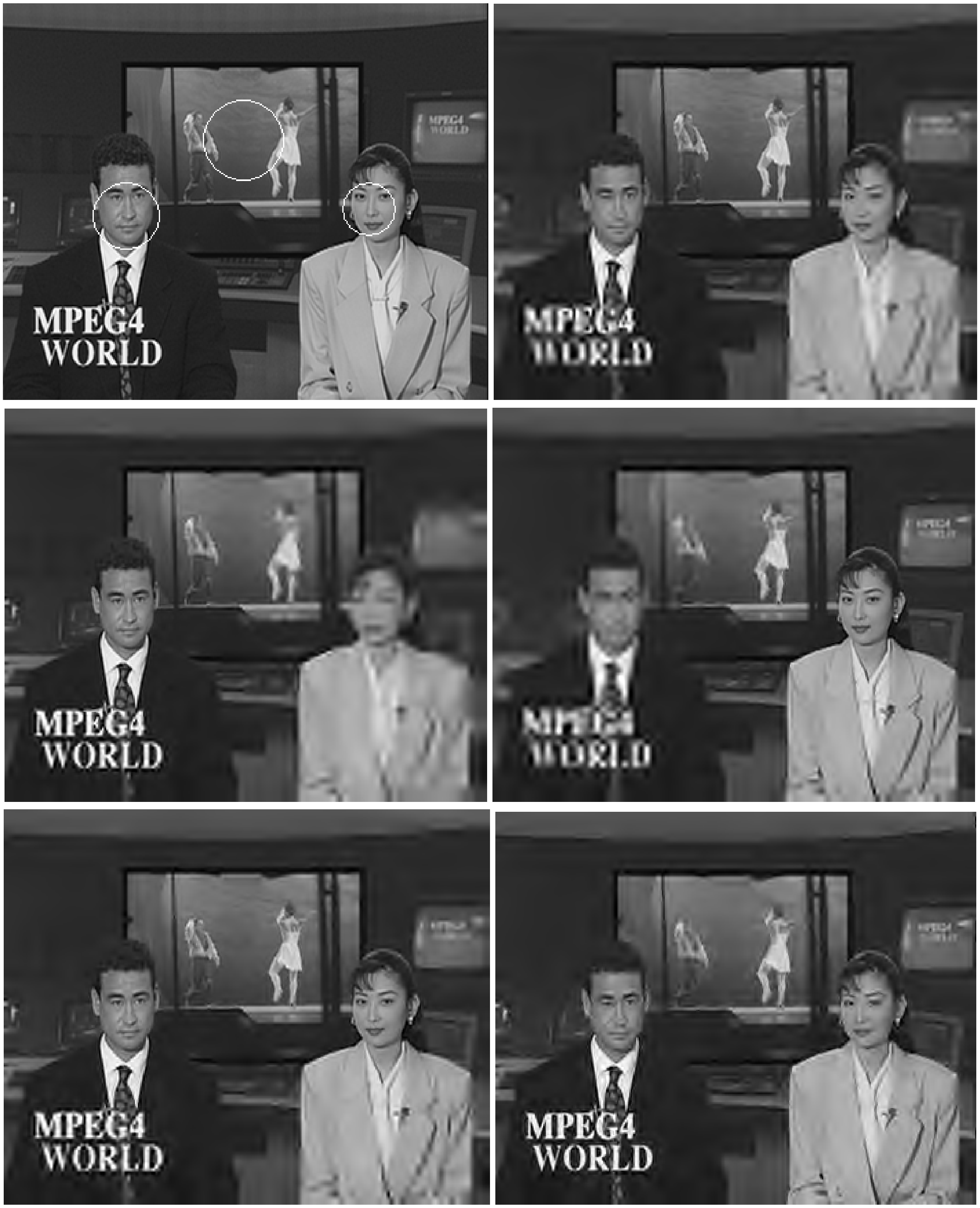

MPEG4
WORLD

Fig. 7 0.25bpp (CR=32:1) “News” image compression result comparison. Top-left: Original image with foveated ROIs indicated; Top-right: EFIC with the upper ROI only; Mid-left: EFIC with the lower left ROI only; Mid-right: EFIC with the lower right ROI only; Bottom-left: EFIC with all the three ROIs; Bottom-right: SPIHT uniform resolution compression.