Embedded Foveation Image Coding

Zhou Wang, *Student Member*, *IEEE*, and Alan C. Bovik, *Fellow*, *IEEE*

Abstract

The human visual system (HVS) is highly space-variant 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. By taking advantage of this fact, it is possible to remove considerable high frequency information redundancy from the peripheral regions and still reconstruct a perceptually good quality image. Great success has been obtained recently by a class of embedded wavelet image coding algorithms, such as the embedded zerotree wavelet (EZW) and the set partitioning in hierarchical trees (SPIHT) algorithms. Embedded wavelet coding not only provides very good compression performance, but also has the property that the bitstream can be truncated at any point and still be decoded to recreate a reasonably good quality image.

In this paper, we propose an embedded foveation image coding (EFIC) algorithm, which orders the encoded bitstream to optimize foveated visual quality at arbitrary bit rates. A foveation-based image quality metric, namely foveated wavelet image quality index (FWQI), plays an important role in the EFIC system. We also developed a modified SPIHT algorithm to improve the coding efficiency. Experiments show that EFIC integrates foveation filtering with foveated image coding and demonstrates very good coding performance and scalability in terms of foveated image quality measurement.

Index Terms – image coding, embedded coding, human visual system, foveation, foveation filtering, foveated image coding, wavelet, progressive transmission, scalable coding, image quality measurement.

EDICS: *1-STIL*

I. INTRODUCTION

The photoreceptors (cones and rods) and ganglion cells are non-uniformly distributed in the retina in the human eye [1, 2]. The density of cone receptors and ganglion cells play important roles in determining the ability of our eyes to resolve what we see. Spatially, the resolution, or sampling density, has the highest value at the point of the fovea and drops rapidly away from that point as a function of eccentricity. As a result, when a human observer gazes at a point in a real world image, a variable resolution image is transmitted through the front visual channel into the high level processing units in the human brain. The region around the point of fixation (or foveation point) is projected into the fovea, sampled with the highest density, and perceived with the highest contrast sensitivity. The sampling density and contrast sensitivity decrease dramatically with increasing eccentricity. In conclusion, the human visual system (HVS) is space-variant in sampling, coding, processing and understanding visual information. By contrast, traditional digital computer vision systems represent images on rectangular uniformly sampled lattices, which have the advantages of simple acquisition, storage, indexing and computation. Nowadays, most digital images and videos are stored, processed, transmitted and displayed in rectangular matrix format, where each entry represents one sampling point.

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 distances can be discovered. There has been growing recent interest in research work on foveated image processing [3-19]. One research direction is *foveation filtering*, which aims to 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. An example is given in Fig. 1, where Fig. 1(a) is the original 'Lena' image and Fig. 1(b) is a foveated version of that image. If attention is focussed at the central foveation point, then the foveated and the original images have the same appearance (depending on the viewing distance). Another research focus is *foveated image and video compression*, which takes advantage of the foveation feature to improve image and video coding efficiency. Perfect foveation with continuously varying resolution turns out to be a difficult theoretical as well as implementation problem. In practice, there are several ways to approximate perfect foveation. A

very commonly used approach is the logmap [3] superpixel method. In [3-6], local pixel groups are averaged and mapped into superpixels, whose sizes are determined by the retinal sampling density. In [7], a multistage superpixel approach is introduced, where a progressive transmission scheme is implemented by using variable sizes of superpixels in each stage. The general idea of logmap can also be employed by a non-uniform sampling scheme. In [8-9], uniform grid images are resampled with a variable resolution that matches the human retina. B-Spline interpolation is then used to reconstruct the foveated images. In [10], a pyramid structure is suggested to foveate images and videos. This structure delivers the possibility of real-time foveated video coding and transmission. In [11-14], foveated images are obtained by applying a foveation filter, which consists of a bank of low pass filters having variable cut-off frequencies. MPEG/H.263 video coding is applied to foveation filtered video sequences. Very good coding performance was obtained because a large amount of visually redundant high frequency information is removed during the foveation filtering processes. In [11-14], the quality of the foveated images are measured uniformly in retinal coordinates. This is equivalent to a non-uniformly weighted measurement, namely foveated peak-to-signal-ratio (F-PSNR), in the original coordinate. The video coding scheme is then designed to optimize F-PSNR. Multiresolution decomposition provides us with a convenient way to simultaneously examine localized spatial as well as frequency information. In [15], a Laplacian pyramid architecture was introduced for image coding. This architecture was utilized in a pyramid vision machine to build a smart sensing system [16]. In [17-19], a wavelet-based foveation method was proposed which applies a non-uniform weighting model in the wavelet transform domain. A progressive transmission method is also suggested, where the image information to be transmitted is ordered according to the weighting of the wavelet coefficients.

Wavelet based image coding algorithms have achieved great success in recent years. The success relies on the energy compaction feature of the discrete wavelet transforms (DWTs) and the efficient organization, quantization, and encoding of the wavelet coefficients. A class of embedded coding algorithms has recently received great attention. The most well-known algorithms are Shapiro's embedded zerotree wavelet (EZW) algorithm [20], and Said and Pearlman's set partitioning in hierarchical trees (SPIHT) algorithm [21], which is an improved implementation of the EZW idea. Embedded wavelet image coding algorithms not only provide very good coding performance, but also

have the property that the bitstream can be truncated at any point and still be decoded to yield a reasonably good quality image. This is a very attractive property that allows for scalable encoding and progressive transmission.

Basically, the EZW and SPIHT encoders try to order the output bitstream, such that those bits with greater contribution to the mean squared error (MSE) between the original and the compressed images are encoded and transmitted first. In other words, the progressive encoding scheme intends to minimize the MSE at any bit rate. HVS features are not considered. However, it is widely accepted that human visual perception distortion does not correlate very well with the MSE [22]. In recent years, researchers have attempted to measure visual image quality in the wavelet transform domain [23-25] and to incorporate HVS-based image quality models with wavelet image coding [26-28]. In [24, 26, 27], the error sensitivity of the wavelet coefficients in different subbands was measured. In [24], the sensitivity measurement results in a model of noise detection thresholds in the wavelet transform domain. The model is merged into a wavelet visible difference predictor in [25]. By employing the perceptual criteria introduced in [26], the authors of [28] proposed a perceptual embedded zerotree image coder. However, their HVS model does not take into account foveation.

The goal of this paper is to design an embedded foveation image coding (EFIC) system, which tries to order the output bitstream, so that those bits with greater contribution to the foveated visual distortion are encoded and transmitted first. In other words, it is designed to optimize foveated visual quality at any bit rate. A foveated image quality metric called foveated wavelet image quality index (FWQI) plays an important role in the system. The bitstream ordering procedure is based on a visual importance weighting model derived from FWQI. EFIC then merges the weighting model into a modified SPIHT encoder, resulting in an efficient foveated image coding system. EFIC can be viewed as an integration of foveation filtering and foveation image coding. This is different from the algorithms in [11-14], where foveation filtering and foveation image/video coding are two separable procedures. The advantages are manifold. First, there is a considerable decrease in computation. Second, EFIC allows scalable encoding and progressive transmission. Third, there is useful tradeoff between bit-rate and the depth of foveation. This can be done simply by truncating the encoded bitstream at any place. Fourth, given enough bandwidth, a high quality uniform resolution image is still attainable. EFIC is also different from the progressive

transmission method proposed in [18]. First, a more sophisticated HVS quality model is used. The model is a joint consideration of multiple factors of the HVS, including the spatial variance of the contrast sensitivity function, the spatial variance of the local visual cut-off frequency, the variance of the human visual sensitivity in different wavelet subbands, and the influence of the viewing distance on the display resolution and the HVS features. Second, the ordering of the transmitted information not only depends on the HVS model's weighting value, but also on the magnitudes of the wavelet coefficients. Third, an efficient embedded coding algorithm is developed especially for the weighted wavelet coefficients to improve the coding efficiency.

II. FOVEATED WAVELET IMAGE QUALITY INDEX (FWQI)

A. Foveation Resolution and Error Sensitivity Models

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 the 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. The variation of the densities of photoreceptors and ganglion cells with eccentricity is shown in Fig. 2. These density distributions play important roles in determining the resolution ability of the human eye.

Psychological experiments have been conducted to measure the contrast sensitivity as a function of retinal eccentricity [10, 29-31]. In [10], a model that fits the experimental data was given by

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

(1)

where f is the spatial frequency (cycles/degree), e is the retinal eccentricity (degrees), CT_0 is the minimal contrast threshold, α is the spatial frequency decay constant, e_2 is the half-resolution eccentricity constant, and $CT(f, e)$ is the visible contrast threshold as a function of f and e. The best fitting parameter values given in [10] are $\alpha = 0.106$, $e_2 = 2.3$, and $CT_0 = 1/64$. It was also reported in [10] that the same values of a

and e_2 provide a good fit to the data in [31] with $CT_0 = 1/75$, and an adequate fit to the data in [30] with $CT_0 = 1/76$, respectively. We adopt this model in our system and use $\alpha = 0.106$, $e_2 = 2.3$, and $CT_0 = 1/64$. The contrast sensitivity is defined as the inverse of the contrast threshold. That is,

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

For a given eccentricity e, equation (1) can be used to find its critical frequency or so called cutoff frequency f_c in the sense that any higher frequency component beyond it is invisible. f_c can be obtained by setting CT to 1.0 (the maximum possible contrast) and solving for e

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

To apply these models to digital images, we need to calculate the eccentricity for any given point $x = (x_1, x_2)^T$ (pixels) in the image. Fig. 3 illustrates a typical viewing geometry. For simplicity, we assume 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 $x^f = (x_1^f, x_2^f)^T$ (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 x to x_f is then $u = \dfrac{d(x)}{N}$, where $d(x) = \|x - x^f\|_2 = \sqrt{(x_1 - x_1^f)^2 + (x_2 - x_2^f)^2}$ (measured in pixels). The eccentricity is given by

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

In Fig. 4, we show 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 given by

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

This approximation is equivalent to that given in [24]. 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 the display resolution:

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

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

$$f_m(\boldsymbol{x}) = \min(f_c, f_d) = \min\left(\frac{e_2 \ln(1/CT_0)}{\alpha\left[e_2 + \tan^{-1}(d(\boldsymbol{x})/Nv)\right]}, \frac{\pi N v}{360} \right) \tag{7}$$

Fig. 5 shows $f_m(\boldsymbol{x})$ for $v = 1, 3, 6$ and 10, respectively. It turns out that 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. 4. Finally, we define the foveation-based error sensitivity for given viewing distance v, frequency f and location $\boldsymbol{x}$ as:

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

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

B. Error Sensitivity in the DWT Domain

The DWT has proved to be a powerful tool for image processing and coding [32-36, 20, 21]. In the one-dimensional 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 several 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 several times. The typical DWT structure is given by Fig.6 (a).

Recently, the 9/7 biorthogonal filters [37] have been widely adopted for DWT-based image compression algorithms. We also use the 9/7 filters in this paper. Readers can refer to [32-37] for more details regarding the basis of wavelet transforms and wavelet-based image processing and coding.

The wavelet coefficients at different subbands and locations supply information of variable perceptual importance to the HVS. In order to develop a good wavelet-based image coding algorithm that considers HVS features, we need to measure the visual importance of the wavelet coefficients. In [24], psychovisual experiments were conducted to measure the visual sensitivity in wavelet decompositions. Noise was added to the wavelet coefficients of a blank image with uniform mid-gray level. After the inverse wavelet transform, the noise threshold in the spatial domain was tested. A model that provided a reasonable fit to the experimental data is [24]:

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

where Y is the visually detectable noise threshold, θ is an orientation index, representing LL, LH, HH, and HL subbands, respectively, and f is the spatial frequency measured in cycles/degree. It is determined by the display resolution r and the wavelet decomposition level λ [24]:

$$f = r2^{-\lambda} \tag{10}$$

The parameters in (9) are tuned to fit the experimental data. For gray scale models, a is 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 detection thresholds for the wavelet coefficients can be calculated by:

$$T_{\lambda,\theta} = \frac{Y_{\lambda,\theta}}{A_{\lambda,\theta}} = \frac{a10^{k(\log(2^\lambda f_0 g_\theta / r)^2}}{A_{\lambda,\theta}} \tag{11}$$

where $A_{\lambda,\theta}$ is the basis function amplitude given in [24]. It is typical to define the error sensitivity as the inverse of the error detection threshold. Therefore, we define the error sensitivity in subband (λ,θ) as:

$$S_w(\lambda,\theta) = \frac{1}{T_{\lambda,\theta}} \tag{12}$$

For a typical viewing distance $v = 3$, the value of S_w for different decomposition levels and orientations

are given in TABLE I.

TABLE I. $S_w(\lambda, \theta)$ FOR THE SIX-LEVEL DWT FOR $v = 3$ AND $N = 512$

Orientation	Wavelet Decomposition Level λ					
	1	2	3	4	5	6
LL	0.3842	0.3818	0.2931	0.1804	0.0905	0.0372
HL	0.2700	0.3326	0.3019	0.2129	0.1207	0.0558
HH	0.1316	0.2138	0.2442	0.2098	0.1430	0.0791
LH	0.2700	0.3326	0.3019	0.2129	0.1207	0.0558

C. Foveation-based Error Sensitivity and Quality Metric in DWT Domain

In order to apply the foveation-based error sensitivity model as equation (8) to the DWT domain, we

first need to calculate the corresponding foveation point in each wavelet subband. For the decomposition

structure given in Fig. 6(a), we have:

$$\text{LL: } \boldsymbol{x}_{\lambda,\theta}^{f} = \left(\frac{x_1^f}{2^\lambda}, \frac{x_1^f}{2^\lambda} \right)^T ; \qquad\qquad \text{LH: } \boldsymbol{x}_{\lambda,\theta}^{f} = \left(\frac{x_1^f + N}{2^\lambda}, \frac{x_1^f}{2^\lambda} \right)^T ;$$

$$\text{HL: } \boldsymbol{x}_{\lambda,\theta}^{f} = \left(\frac{x_1^f}{2^\lambda}, \frac{x_1^f + N}{2^\lambda} \right)^T ; \qquad \text{HH: } \boldsymbol{x}_{\lambda,\theta}^{f} = \left(\frac{x_1^f + N}{2^\lambda}, \frac{x_1^f + N}{2^\lambda} \right)^T . \qquad (13)$$

Next, for a given wavelet coefficient at position $\boldsymbol{x} \in \boldsymbol{B}_{\lambda,\theta}$, where $\boldsymbol{B}_{\lambda,\theta}$ denotes the set of wavelet

coefficient positions residing in subband (λ, θ), its equivalent distance from the foveation point in the

spatial domain is given by

$$d_{\lambda,\theta}(\boldsymbol{x}) = 2^\lambda \left\| \boldsymbol{x} - \boldsymbol{x}_{\lambda,\theta}^{f} \right\|_2 \qquad \text{for } \boldsymbol{x} \in \boldsymbol{B}_{\lambda,\theta} \qquad (14)$$

With this equivalent distance, and also considering (10), we have

$$S_f(v, f, \boldsymbol{x}) = S_f(v, r2^{-\lambda}, d_{\lambda,\theta}(\boldsymbol{x})) \qquad \text{for } \boldsymbol{x} \in \boldsymbol{B}_{\lambda,\theta} \qquad (15)$$

Finally, a foveation-based error sensitivity model in the DWT domain is determined by a combined

consideration of (12) and (15):

$$S(v, \boldsymbol{x}) = \left[S_w(\lambda, \theta) \right]^{\beta_1} \cdot \left[S_f(v, r2^{-\lambda}, d_{\lambda,\theta}(\boldsymbol{x})) \right]^{\beta_2} \quad \text{for } \boldsymbol{x} \in \boldsymbol{B}_{\lambda,\theta} \qquad (16)$$

where β_1 and β_2 are parameters used to control the magnitudes of S_w and S_f, respectively. In this paper, we use $\beta_1 = 1$ and $\beta_2 = 2.5$. The construction of $S(v, x)$ can be viewed as two stages in cascade. In the first stage, each wavelet subband is assigned a uniform base importance value according to $S_w(\lambda, \theta)$. In the second stage, non-uniform weights developed from $S_f(v, r2^{-\lambda}, d_{\lambda,\theta}(x))$ are applied to the subbands, resulting in a space-variant error sensitivity mask in the DWT domain. In Fig. 7, we show the error sensitivity masks for viewing distance $v = 1, 3, 6$ and 10, respectively.

For the evaluation of image quality, instead of using the traditional error summation methods, we designed a new quality index [38] by modeling any signal distortion as a combination of three factors: loss of correlation, mean distortion and variance distortion. For any 2-D signal, the measurement results are a 2-D quality map as well as an overall quality index. Readers can refer to [38] and *http://anchovy.ece.utexas.edu/~zwang/research/quality_index/demo.html* for more details and demonstrative images of the new quality index. In this paper, we adapt the index into the DWT domain and define the foveated wavelet image quality index (FWQI) as:

$$FWQI = \frac{\sum_{n=1}^{M} S(v, x_n) \cdot |c(x_n)| Q(x_n)}{\sum_{n=1}^{M} S(v, x_n) \cdot |c(x_n)|}, \tag{17}$$

where M is the number of the wavelet coefficients, $c(x_n)$ is the wavelet coefficient of the original image at location x_n, and $Q(x_n)$ is the quality value at location x_n in the quality index map. Since $S(v, x_n)$ varies with v, FWQI of a test image is a function of v, instead of a single value.

D. Multiple Foveation Points and Regions

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 $i = 1, 2, \cdots, P$. The overall error sensitivity should be given by the maximum of them:

$$S(v, \boldsymbol{x}) = \max_{i=1\cdots P}(S_i(v, \boldsymbol{x})). \tag{18}$$

In practice, it is not necessary to compute each of the $S_i(v, \boldsymbol{x})$. 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

$$S(v, \boldsymbol{x}) = S_j(v, \boldsymbol{x}), \qquad \text{where } j \in \operatorname*{arg\,min}_{i \in \{1, 2, \cdots, P\}} \left\{ \left\| \boldsymbol{x} - \boldsymbol{x}_i^f \right\|_2 \right\}. \tag{19}$$

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

III. EMBEDDED FOVEATION IMAGE CODING (EFIC)

A. Review of Embedded Wavelet Image Coding Methods

The main objective in embedded wavelet image coding is to choose the most important wavelet coefficients to be encoded and transmitted first. The importance of a coefficient in EZW and SPIHT depends on its contribution to the MSE distortion. The coefficients with larger magnitudes are more important. The strategy is ordering the coefficients by magnitude and transmitting the most significant bits first. Assume that the wavelet coefficients have been ordered according to the minimum number of bits required for its magnitude binary representation. The schematic binary representation is shown in Fig. 8(a) [21]. The most effective order for progressive transmission is to sequentially send the bits in each row, as indicated by the arrows. In order for the decoder to understand the meaning of the bits, we also need to encode and transmit the coordinates of the wavelet coefficients along with the magnitude bits. It has been observed that the wavelet coefficients which are less significant have structural similarity across the wavelet subbands in the same spatial orientation. The zerotree structure in EZW and the spatial orientation tree structure in SPIHT capture this structural similarity very effectively. Fig. 6(b) shows the spatial orientation tree used by SPIHT.

In EZW or SPIHT encoder, the wavelet coefficients are scanned multiple times. Each time consists of a sorting pass and a refinement pass. The sorting pass selects the significant coefficients and encodes the spatial orientation tree structure. A coefficient is significant if its magnitude is larger than a threshold value, which decreases by a factor of 2 for each successive sorting pass. The refinement pass outputs one bit for each selected coefficient, as indicated by the arrows in Fig. 8(a). An entropy coder can be used to further compress the output bitstream. SPIHT performs better than EZW in terms of reconstructed image quality. The embedded coding of EFIC is developed based on a modified version of SPIHT.

B. The EFIC System

The proposed EFIC system is depicted in Fig. 9. We first apply the wavelet transform to the original image. We assume we already know the foveation points and regions, which are used to compute an error sensitivity-based importance weighting mask. The wavelet coefficients are then weighted using the weighting mask. Next, we encode the weighted wavelet coefficients using a modified SPIHT encoder. The output bitstream of the modified SPIHT encoder, together with the foveation parameters, is transmitted to the communication network. At the receiver side, the weighted wavelet coefficients are obtained by applying the modified SPIHT decoding. The importance weighting mask is then calculated in exactly the same way as at the sender side. Finally, the inverse weighting and inverse wavelet transform are applied to obtain the reconstructed image. Between the sender, the communication network and the receiver, it is possible to exchange information about network conditions and user requirements. Such feedback information can be used to control the encoding bit rate and foveation points. The decoder can also truncate the received bitstream to obtain any bit rate image below the encoder bit rate.

There are two key techniques in the EFIC system. One is the calculation of the importance weighting mask. The other is the modified SPIHT algorithm. We will discuss them in the next subsections.

C. Importance Weight Calculation

The purpose of the importance weighting mask is to help the encoder to order the output bitstream, so that bits with greater contribution to the foveated visual quality are encoded and transmitted earlier. Basically, the weight $W(x_n)$ assigned to a wavelet coefficient x_n must be consistent with the foveation-

based error sensitivity model given in (16). Therefore, the desired solution is

$$W(\boldsymbol{x}_n) = K \cdot S(v, \boldsymbol{x}_n),$$ (20)

where K can be any constant value except for 0. In this solution, the viewing distance v must be known to us. However, in many practical applications, v is not available to the encoder. One solution to this problem is to assume a fixed viewing distance. In this paper, we solve it by assuming a probability distribution of viewing distance instead of a fixed one. The probability model is given by:

$$p(v) = \frac{1}{\sqrt{2\pi}\sigma v} \exp\left(-\frac{(\ln v - \mu)^2}{2\sigma^2}\right) \quad for \ v \in (0, \infty)$$ (21)

The distribution curve is shown in Fig. 10, where σ and μ are tuned to $\sigma = 0.4$ and $\mu = 1.2586$, respectively, so that the maximum possible viewing distance is $v = 3$ and the typical viewing distance is between 1.5 and 6 times of the image width. With this distribution, the importance weight of $\boldsymbol{x}_n$ is

$$W(\boldsymbol{x}_n) = \int_{0^+}^{\infty} p(v)S(v, \boldsymbol{x}_n)dv$$ (22)

Fig. 11 shows the resulting importance weighting mask.

D. Modified SPIHT

It is possible to implement an embedded foveation coding system with the original SPIHT algorithm. However, there arises a problem that makes the SPIHT algorithm inefficient. Note that in our system, the input coefficients to the SPIHT part are the weighted wavelet coefficients instead of the original ones. The weighted coefficients have a much larger dynamic range compared to that of the original coefficients. A schematic binary representation of a list of magnitude-ordered weighted wavelet coefficients is shown in Fig. 8(b). Comparing this with Fig. 8(a), we see that the change of dynamic range leads to an increase in the number of times the wavelet coefficients are scanned. This makes SPIHT encoding less efficient in two aspects. First, since we need to encode the spatial orientation tree structure with every scan, an increase in the number of times an image is scanned implies an increase in the wastage of bits and an increase of time for scanning and computation. Second, we are encoding the significant coefficients with more bits as we increase the number of scans because we add one more refinement bit to each of them during each scan. Consequently, a modified SPIHT algorithm is needed to overcome this problem. We

solve this problem in two ways.

First, in the sorting pass, we do not scan all the wavelet coefficients in the first few scans. Suppose the maximum absolute value of the unweighted wavelet coefficients is $\max_{x}\{c(x)\}$, then the largest possible absolute value of the weighted wavelet coefficient at location x is $\max_{x}\{c(x)\} \cdot W(x)$. In the first few scans, it might be lower than the significance test threshold in the SPIHT algorithm. The threshold at the n-th scan is given by $T_n = \left\lfloor \log_2(\max_{x}\{c(x)W(x)\}) \right\rfloor / 2^{n-1}$. Our strategy is that before each scan, we perform the following test: $\max_{x}\{c(x)\} \cdot W(x) \overset{?}{\geq} T_n$. Only those x that satisfy this condition are scanned. The wavelet coefficients that are scanned in the first four times are given in Fig. 12. It can be seen that only a small subset of the coefficients need to be scanned during the first few sorting passes.

Second, in the refinement pass, it is not necessary to encode any of the weighted coefficients using too many bits, as in Fig. 8(b). This is because during the inverse weighting procedure, the coefficients will be scaled back to values on the order of their original values. The less significant bits (such as the 10-th or even less significant bits) do not have much contribution to the overall image quality, therefore can be removed. In EFIC, we limit the maximum number of bits for each coefficient. This is shown in Fig. 8(c). The refinement bits are sent sequentially in the order indicated by the arrows in Fig. 8(c). Actually, any wavelet coefficient that receives the upper limit of refinement bits can be removed from the list of significant pixels of the SPIHT algorithm.

E. Experimental Results

We test the EFIC algorithm using 8 bits/pixel (bpp) gray scale images and compare it with the SPIHT algorithm. Fig. 13 shows the 512×512 "Zelda" image encoded with both SPIHT and EFIC algorithms. At a very low bit-rate of 0.015625bpp with compression ratio (CR) equaling 512:1, the mouth, nose and eye regions are hardly recognizable in the SPIHT coded image, whereas those regions in the EFIC coded image exhibit some detailed visual information. At a low bit-rate of 0.03125bpp (CR=256:1), SPIHT still decodes a very blurred image, while EFIC begins to give acceptable quality over the face region. Increasing the bit rate to 0.0625bpp (CR=128:1) and 0.125bpp (CR=64:1), the visual quality of the EFIC

coded images is still superior to the SPIHT coded images. When the bit rate of 0.25bpp (CR=32:1) is reached, the EFIC coded image approaches uniform resolution, and the decoded SPIHT and EFIC images are almost indistinguishable. The EFIC decoding can also be viewed as a foveation filtering procedure with decreasing foveation depth. Notice 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 coefficients. 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 consistent with 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. 13 demonstrate this very well.

Fig. 14 shows the FWQI comparisons of the EFIC and SPIHT compressed "Zelda" images at 0.015265, 0.0625 and 0.25bpp. FWQI is given as a function of the viewing distance, instead of just one fixed value. In comparison with SPIHT, significant quality gain is achieved by EFIC through the entire range of viewing distances. This is consistent with the subjective quality shown in Fig. 13. In Fig. 15(a), we show how the FWQI value increases with the bit rate. Fig. 16 shows the EFIC compression of the "Board" image with multiple foveation regions. At low bit rates such as 0.0625bpp and 0.125bpp, EFIC maintains acceptable quality at the foveation regions and blurs the regions of lower interest. Again, a visually high quality uniform resolution image is obtained from the same bit stream with a sufficient bit-rate (0.5bpp). The FWQI results of EFIC compressed "Board" images are given in Fig. 15(b).

In Fig. 17, we compare the 288×352 "News" image compression results with the same bit rate but different foveation region selections. It turns out that with a bit rate of 0.25bpp, uniform resolution SPIHT coding cannot provide an acceptable image, but if the foveation region(s) are known to us, visually satisfactory image quality is still achievable with the EFIC algorithm.

IV. Discussions and Future Work

When we introduce our foveation image coding and processing work to people, one of the most

frequently asked questions 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 [39-41]. 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 [41]. There is little doubt that our foveation-based techniques will be more effective if combined with a very intelligent image analysis, pattern recognition, and image understanding system. Actually, we are now conducting research on visual fixation modeling. Although it is argued that it is always difficult to decide foveation points, we believe that it is feasible to establish a statistical model. 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 then used to track the user-selected object as the foveated region in the image sequence that follows.

In general, the EFIC algorithm and the video coding algorithm developed from it are good for low and variable bit-rate image and video communication applications. Foveation-adaptive scalable coding and foveation-progressive transmission are the key features. There is a lot of future work left to do.

One direct application of EFIC is Internet browsing [18]. There are two significant examples. In the first, the point(s) of fixation is predetermined and one copy of the EFIC encoded bitstream of the high quality image is stored at the server side. During transmission, the client receives and decodes a highly foveated image first, and with the arrival of more bits, the quality of the image is gradually refined. Finally, a high quality, uniform resolution image is achieved. In the second example, the wavelet coefficients, together with a uniform coarse quality version of the image, are stored at the server side without EFIC encoding. The client first sees the coarse version of the image and figures out the point of interest. The selected point is sent back to the server and activates the EFIC encoding. The encoded bitstream is then transmitted to the client with a foveation emphasis at the selected point of interest.

Currently, we are also working on real time foveation-scalable video coding and communication systems over computer networks, where the points and depth of foveation are determined by feedback from the receiver and the network. Feedback from the receiver can be the fixation points, the decoder buffering situation and the data consumption speed. The feedback from the network includes the change of bandwidth, the network congestion situation, and the latency. Given the feedback information, the foveated system at the server side then adaptively varies the foveation points as well as the bit-rate by changing the foveation depth and the frame rate. This is superior to the current systems. Real time implementation is important for a successful foveated video communication system. We are also doing research and development work with digital signal processing chips to implement fast foveation filtering and fast foveation image and video coding.

References

[1]   B. A. Wandell, *Foundations of Vision*, Sinauer Associates, Inc., 1995.

[2]   L. K. Cormack, "Computational models of early human vision," in *Handbook of Image and Video Processing*, Al Bovik, ed., Academic Press, 2000.

[3]   R. S. Wallace, P.-W. Ong, B. B. Bederson, and E. L. Schwartz, "Space variant image processing," *International Journal of Computer Vision*, vol. 13, no. 1, pp. 71-90, 1994.

[4]   P. Kortum, W. S. Geisler, "Implementation of a foveated image coding system for image bandwidth reduction," *Human Vision & electronic Imaging*, *Proc. of SPIE*, vol. 2657, pp. 350-360, 1996.

[5]   C. Bandera, P. Scott, "Foveal machine vision systems," *IEEE International Conference on Systems, Man and Cybernetics*, pp. 596-599, Cambridge, Nov. 1989.

[6]   P. Camacho, F. Arrebola, F. Sandoval, "Shifted fovea multiresolution geometries," *IEEE International Conference on Image Processing*, vol. 1, pp. 307-310, New York, 1996.

[7]   N. Tsumura, C. Endo, H. Haneishi, and Y. Miyake, "Image compression and decomposition based on gazing area," *Human Vision and Electronic Imaging*, *Proc. of SPIE*, vol. 2657, pp.361-367, 1996.

[8]   T. Kuyel, W. Geisler, and J. Ghosh, "Retinally reconstructed images: digital images having a resolution match with the human eyes," *IEEE Trans. System, Man, & Cybernetics, - Part A: Systems and Humans*, vol. 29, no. 2, pp. 235-243, Mar. 1999.

[9]   T. Kuyel, "Foveated models in compression, texture discrimination and classification," *Ph.D. Dissertation*, The University of Texas at Austin, May 1997.

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

[11] S. Lee, M. S. Pattichis and A. C. Bovik, "Foveated video quality assessment," to appear in *IEEE Trans. on Multimedia*, 2001.

[12] S. Lee, M. S. Pattichis and A. C. Bovik, "Optimal rate control low bitrate foveated video coding," to appear in *IEEE Trans. on Image Processing*, 2001.

[13] S. Lee, M. S. Pattichis, and A. C. Bovik, "Rate control for foveated MPEG/H.263 video," *IEEE International Conference on Image Processing*, vol. 2, 1998.

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

[15] P. J. Burt and E. H. Adelson, "The Laplacian algorithm as a compact image code," *IEEE Trans. Communications*, vol. 31, no. 4, pp. 532-540, Apr. 1983.

[16] P. J. Burt, "Smart sensing within a pyramid vision machine," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 1006-1015, Aug. 1988.

[17] E.-C. Chang, and C. Yap, "A wavelet approach to foveating images", *Proceedings of the 13th ACM Symposium on Computational Geometry*, pp. 397-399, 1997.

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

[19] E.-C. Chang, S. Mallat, and C. Yap, "Wavelet foveation," http://www.cs.nyu.edu/visual/, Jan. 1999.

[20] J. M. Shapiro, "Embedded image coding using zerotrees of wavelets coefficients," *IEEE Trans. Signal Procesing*, vol. 41, pp. 3445-3462, Dec. 1993.

[21] 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.

[22] B. Girod, "What's wrong with Mean-Squared Error," *Digital Images and Human Vision*, A. B. Watson, ed., 1993.

[23] Y. K. Lai, and C.-C. J. Kuo, "A Haar wavelet approach to compressed image quality measurement," *Journal of Visual Communication and Image Representation*, vol. 11, pp. 17-40, Mar. 2000.

[24] 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.

[25] A. P. Bradley, "A wavelet visible difference predictor," *IEEE Trans. Image Processing*, vol. 8, no. 5, pp. 717-730, May 1999.

[26]   R. J. Safranek, and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," *IEEE International Conference Acoustics, Speech, and Signal Processing*, pp. 1945-1948, May 1989.

[27]   P. N. Topiwala, "Human vision-based wavelet image coding," *Proceedings of SPIE*, vol. 3164, pp. 260-270, 1997.

[28]   I. Hontsch, L. J. Karam, and R. J. Safranek, "A perceptually tuned embedded zerotree image coder," *IEEE International Conference on Image Processing*, pp. 41-44, Oct. 1997.

[29]   J. G. Robson, and N. Graham, "Probability summation and regional variation in contrast sensitivity across the visual field," *Vision Research*, vol. 21, pp. 409-418, 1981.

[30]   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.

[31]   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.

[32]   S. G. Mallat, "Multifrequency channel decomposition of images and wavelet models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 2091-2110, 1989.

[33]   M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using the wavelet transform," *IEEE Trans. Image Processing*, vol. 1, no. 2, pp. 205-220, 1992.

[34]   Z. Xiong, and K. Ramchandran, "Wavelet image compression," in *Handbook of Image and Video Processing,* Al Bovik, ed., Academic Press, 2000.

[35]   J. Villasenor, b. Belzer, and J. Liao, "Wavelet filter evaluation for efficient image compression," *IEEE Trans. Image Processing*, vol. 4, pp. 1053-1060, 1995.

[36]   M. Vetterli, and J. Kovačević, *Wavelets and Subband Coding*, Prentice Hall PTR, Englewood Cliffs, New Jersey, 1995.

[37]   A. Cohen, I. Daubechies, and J. C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Commun. Pure Appl. Math.*, vol. 45, pp. 485-560, 1992.

[38]   Z. Wang, and A. C. Bovik, "A universal image quality index," submitted to *IEEE Signal Proc. Letters*, 2001.

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

[40]   J. M. Findlay, R. Walker, and R. W. Kentridge, ed., *Eye Movement Research: Mechanisms, Processes and Applications*, Elsevier Science, North-Holland, 1995.

[41]   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.

(a)   (b)

Fig. 1   An example of foveated image. (a) original "Lena" image; (b) foveated "Lena" image.
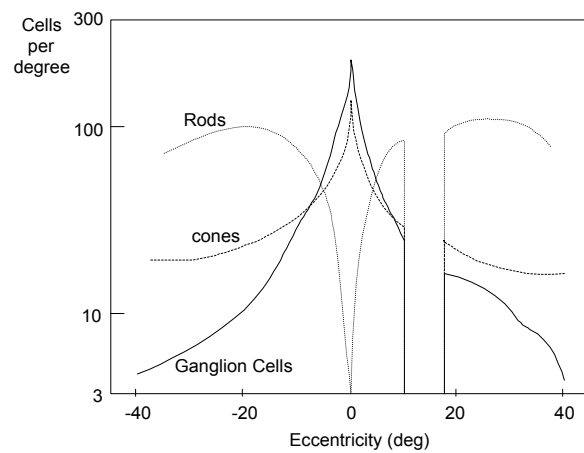
Fig. 2   Photoreceptor and ganglion cell density as a function of eccentricity.
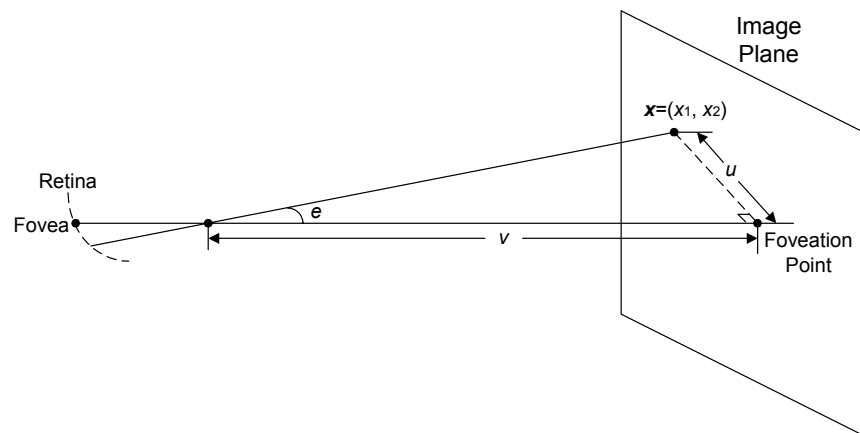
Fig. 3   A typical viewing geometry.
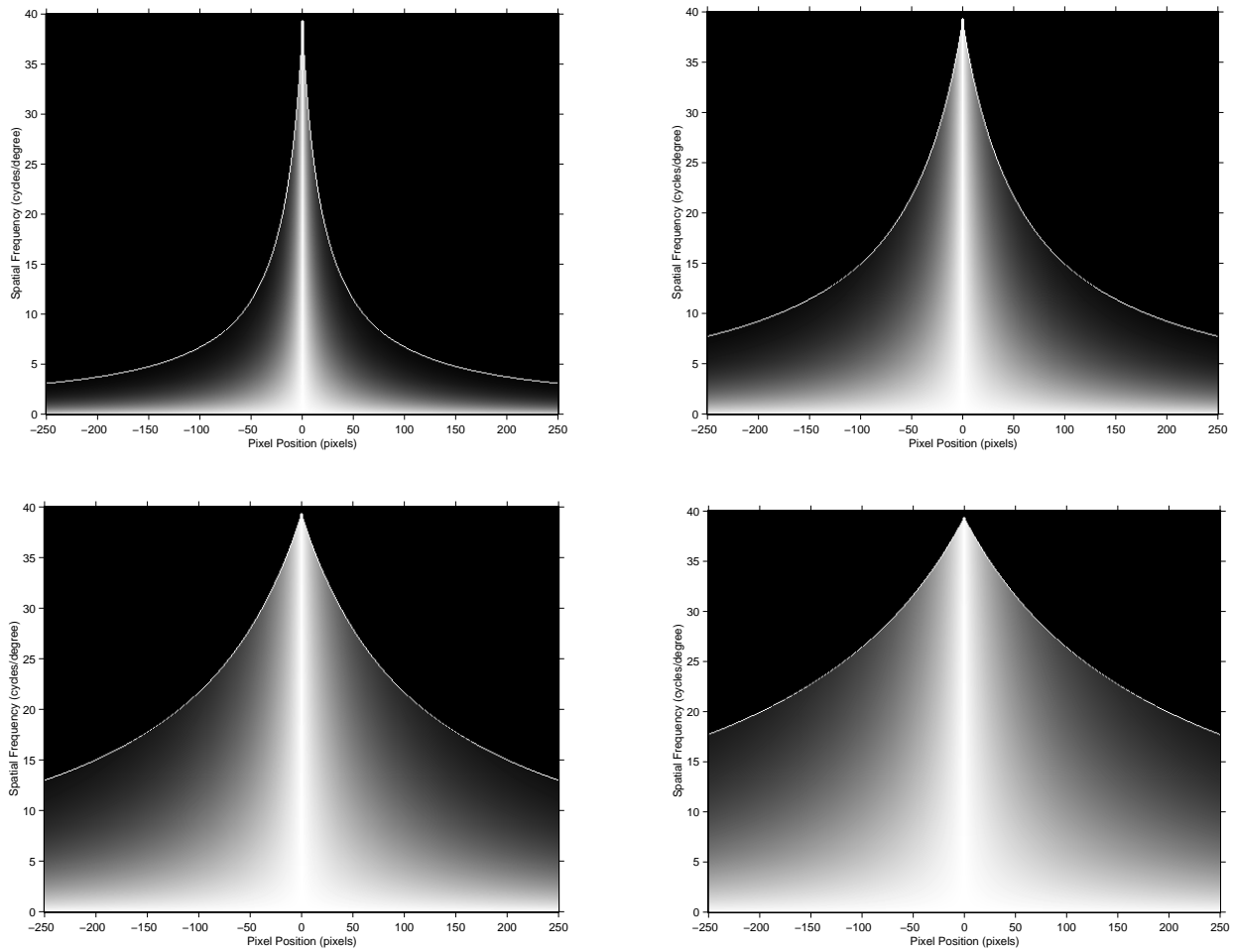
Fig. 4 Normalized contrast sensitivity (Brightness indicates the strength of contrast sensitivity). The top-left, top-right, bottom-left and bottom-right figures are for $N = 512$ and viewing distance $v = 1$, 3, 6 and 10 times of the image width, respectively. The white curves show the cutoff frequency.
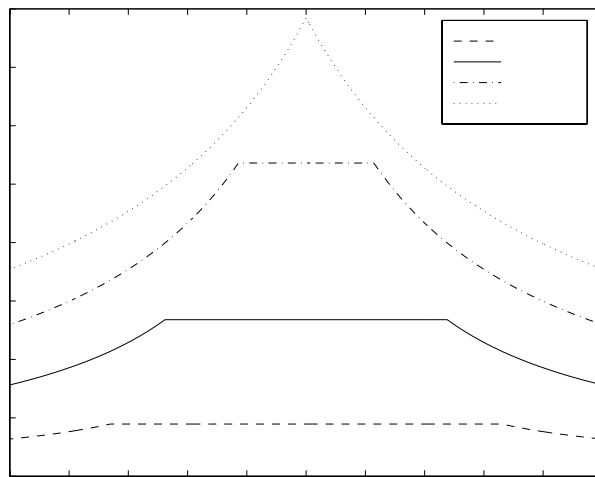
Fig. 5   Nyquist frequency limited cutoff frequency plotted as a function of pixel position in the image.
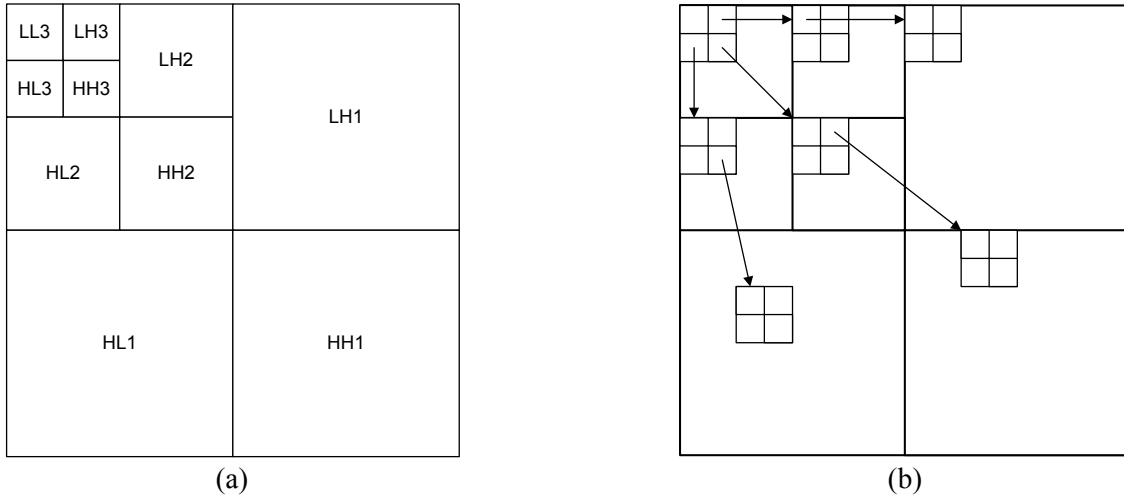
Fig. 6 (a) DWT decomposition structure; (b) Spatial orientation tree in SPIHT algorithm.
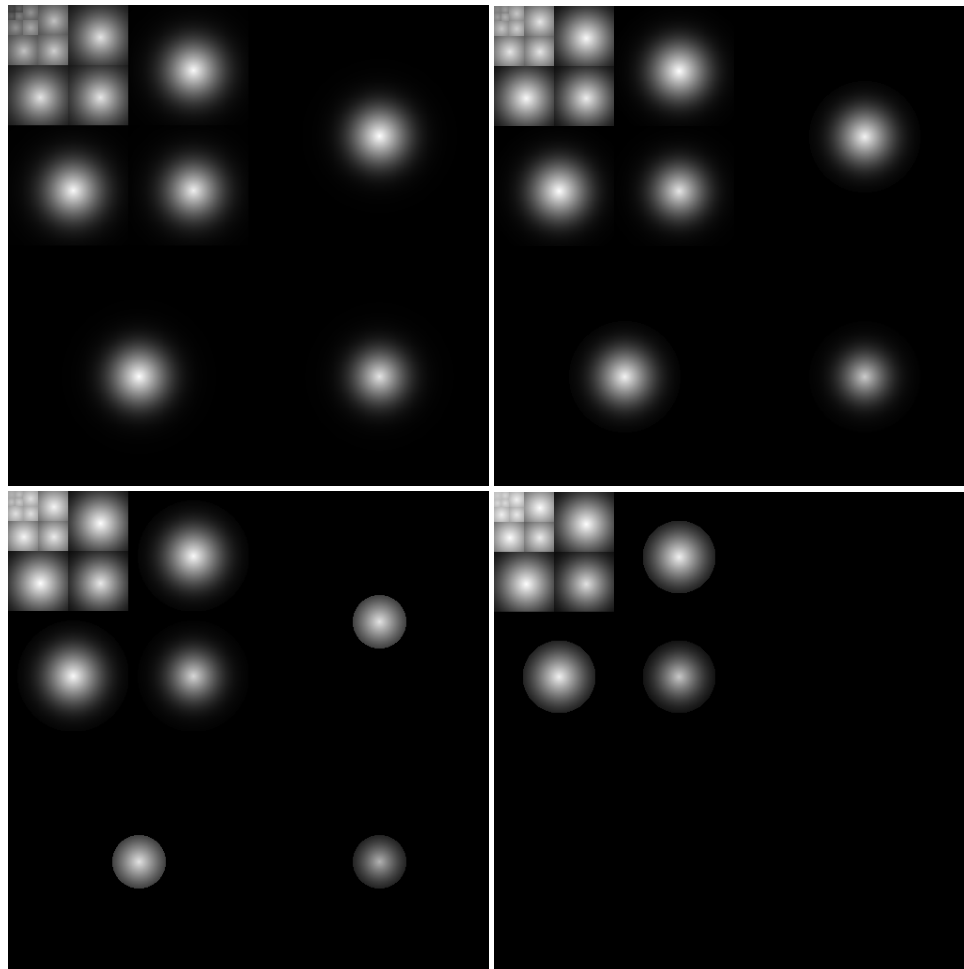
Fig. 7 Foveation-based error sensitivity mask in the DWT domain. The top-left, top-right, bottom-left, and bottom-right figures are for viewing distance $v = 1, 3, 6$ and 10 times of the image width, respectively. (Brightness logarithmically enhanced for display purpose)

(a)

(b)

(c)

Fig. 8 (a) Binary representation of magnitude-ordered wavelet coefficients. (b) Binary representation of magnitude-ordered weighted wavelet coefficients in SPIHT algorithm; (c) Binary representation of magnitude-ordered weighted wavelet coefficients in modified SPIHT algorithm.
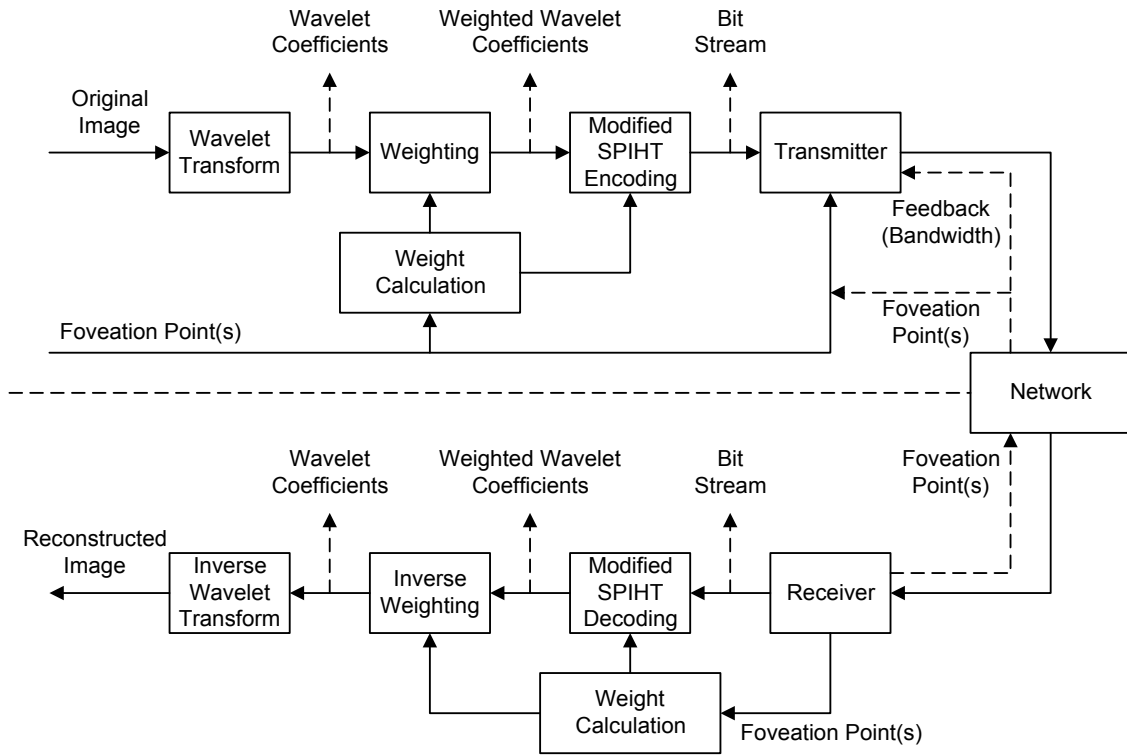
Fig. 9   The EFIC system.
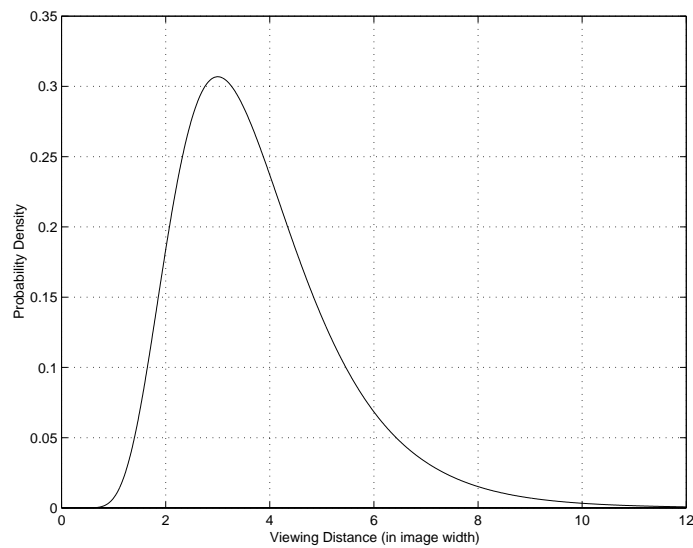
Fig. 10   Probability distribution of viewing distance v.
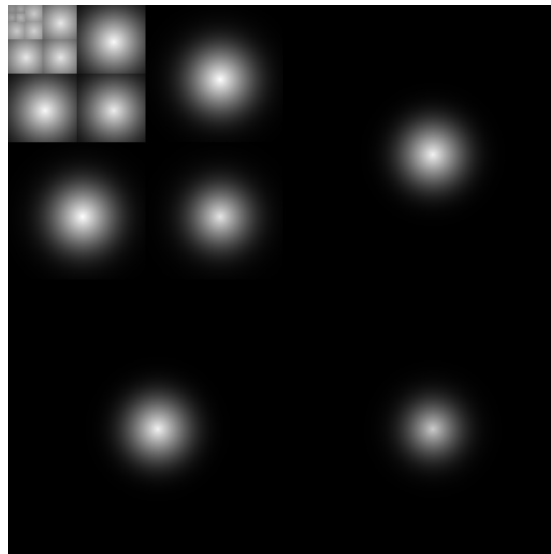
Fig. 11 Importance weighting mask. Brightness indicates the importance of the wavelet coefficients (Brightness logarithmically enhanced for display purpose).
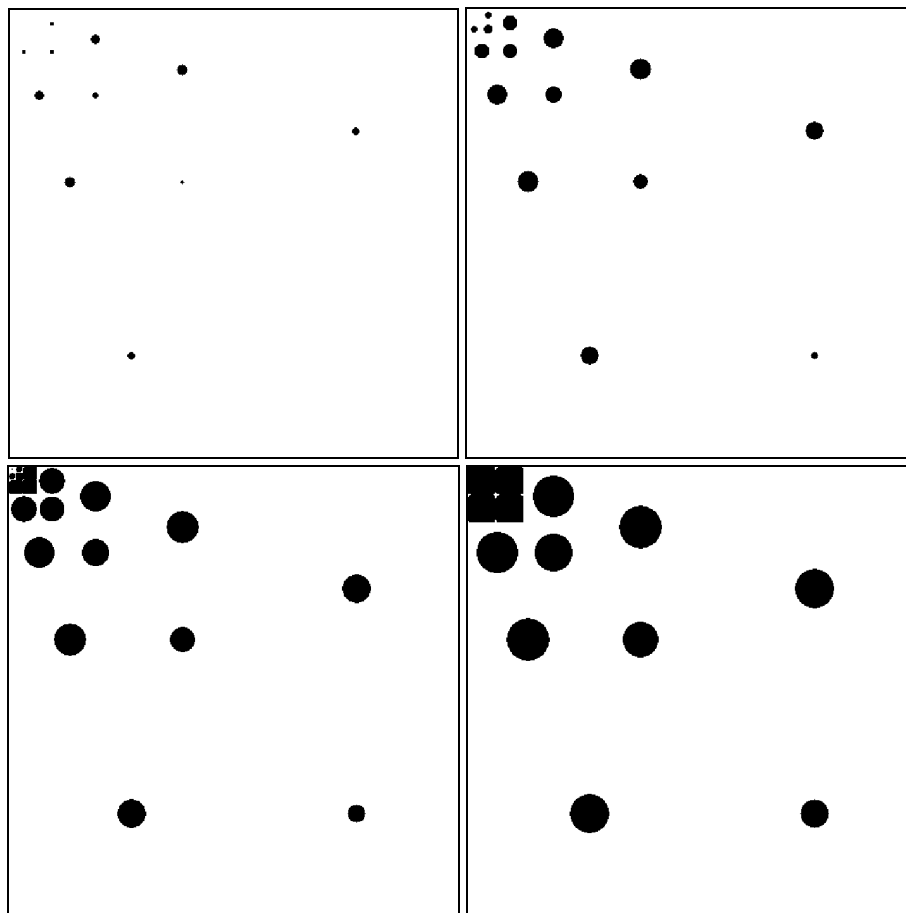
Fig. 12   Wavelet coefficients scanned at the first four times in the modified SPIHT algorithm.

Fig. 13 "Zelda" image compression results. Top-left: Original image with the foveation region indicated; Top-right: DWT domain importance weighting mask for EFIC (Brightness logarithmically enhanced for display purpose); The images of the left column that follow: SPIHT coded images; The images of the right column that

follow: 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.
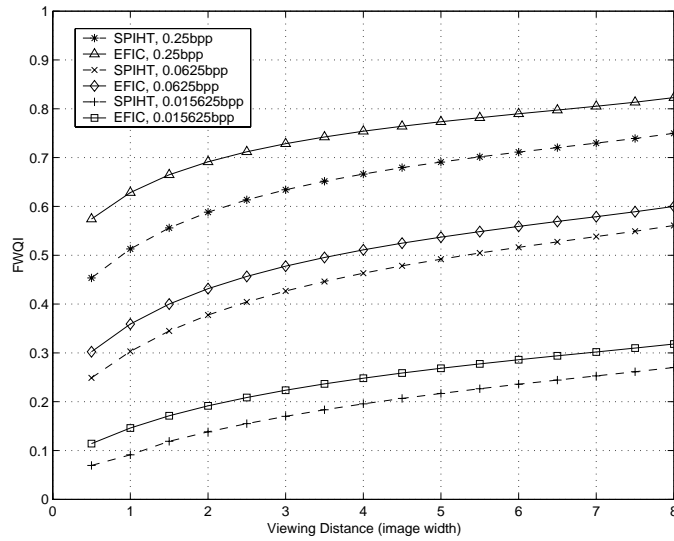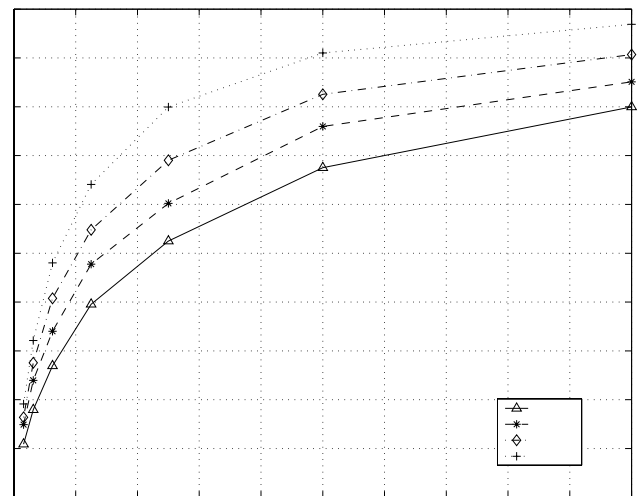

Fig. 14 FWQI comparison of EFIC and SPIHT compressed "Zelda" image at 0.15625bpp, 0.0625bpp and 0.25bpp.
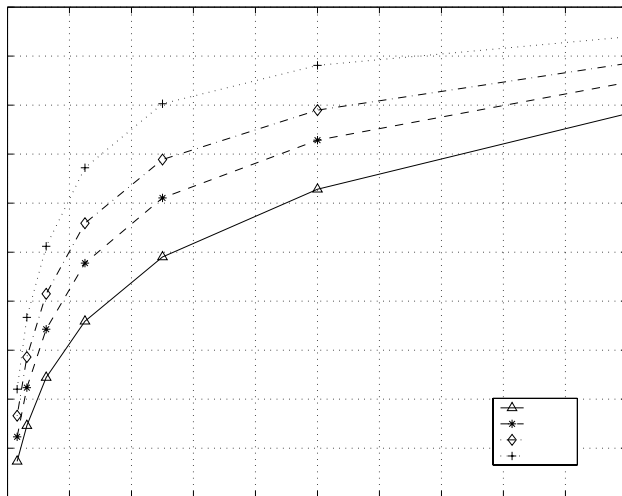
(a)   (b)

Fig. 15 FWQI results of EFIC compressed "Zelda" (a) and "Board" (b) images.
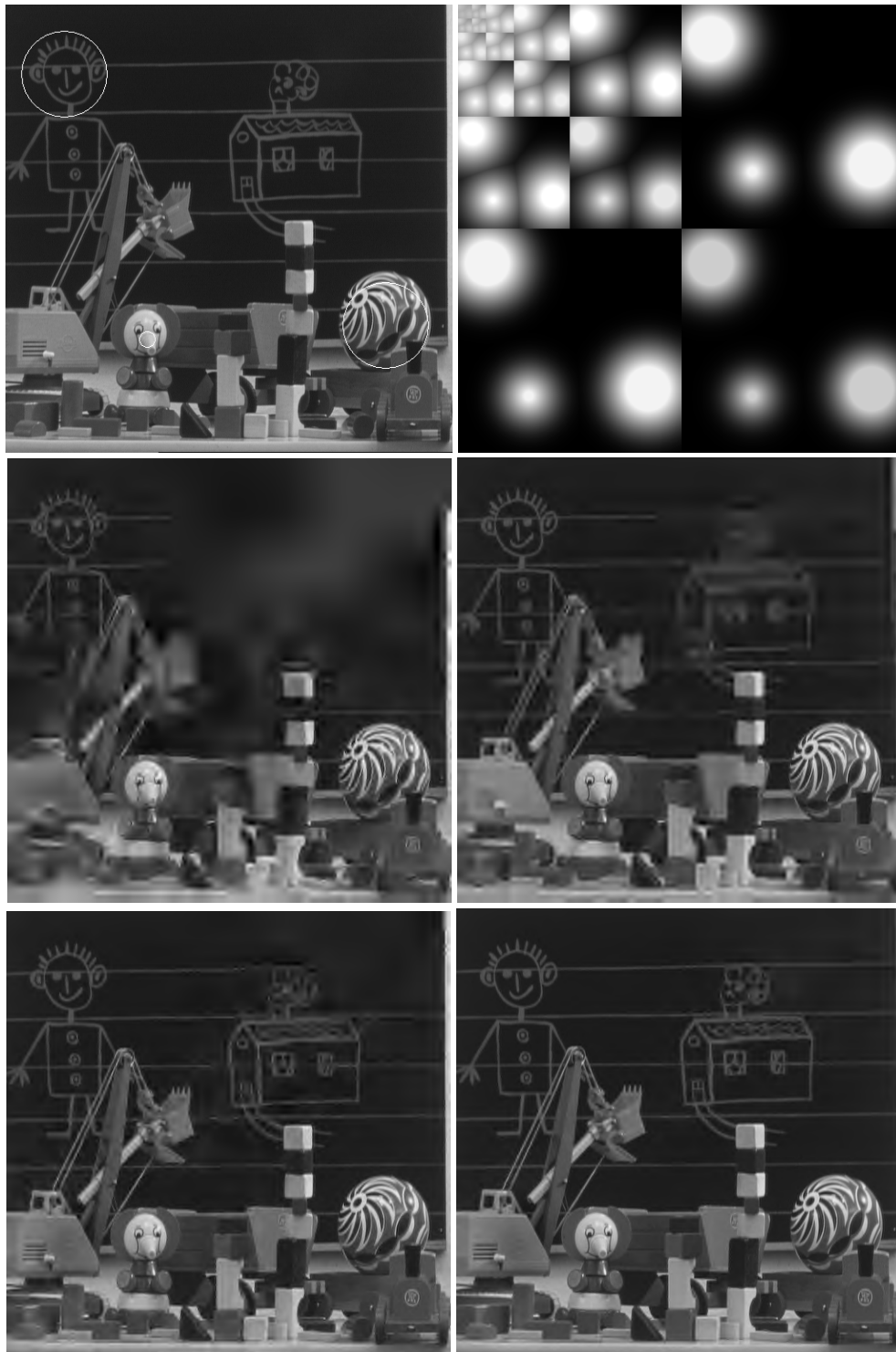
Fig. 16 Top-left: "Board" image with multiple foveation regions indicated; Top-right: DWT domain importance weighting mask (Brightness logarithmically enhanced for display purpose); Mid-left: EFIC compression, 0.0625bpp (CR=128:1); Mid-right: EFIC

compression, 0.125bpp (CR=64:1); Bottom-left: EFIC compression, 0.25bpp (CR=32:1); Bottom-right: EFIC compression, 0.5bpp (CR=16:1).
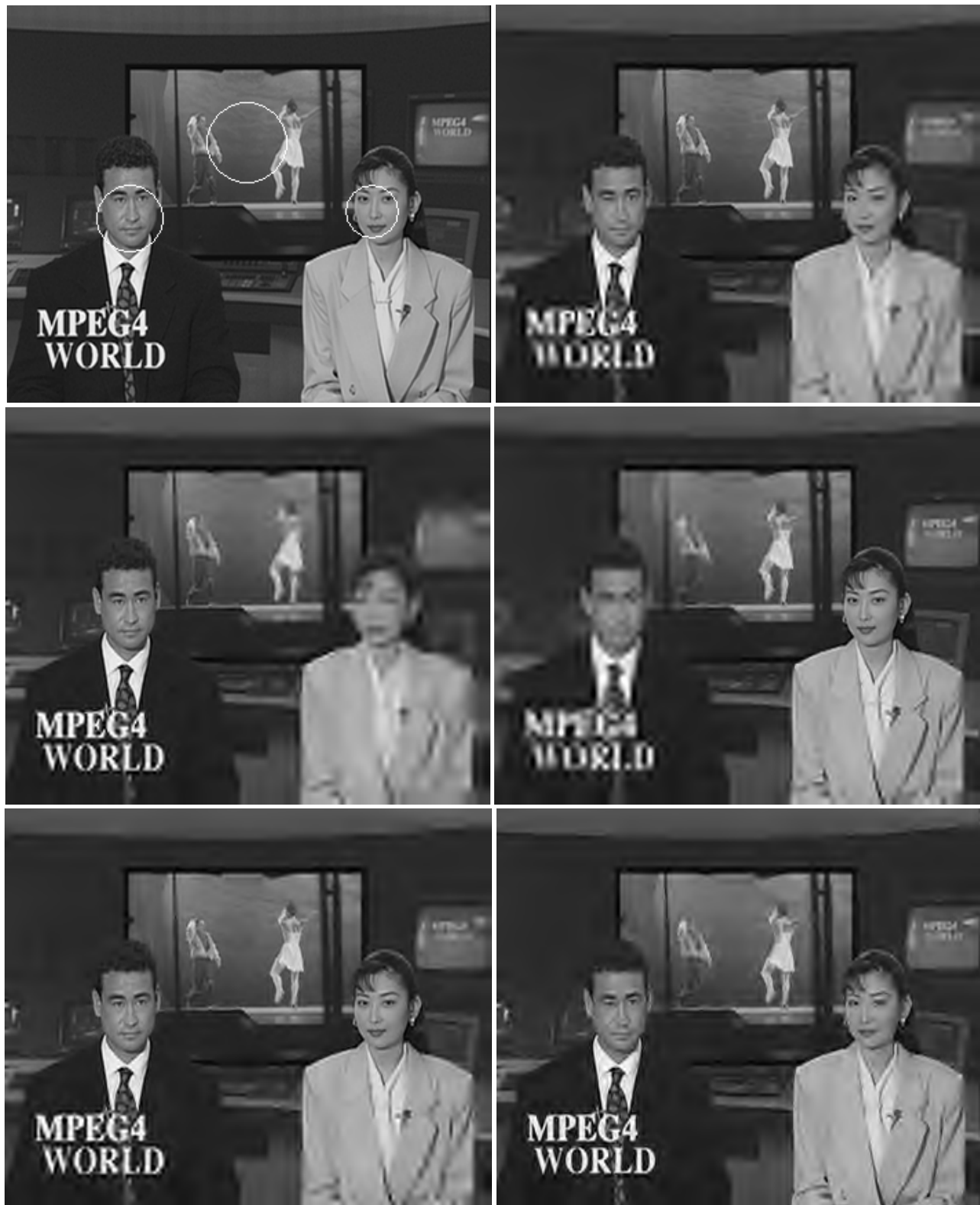
Fig. 17 0.25bpp (CR=32:1) "News" image compression results. Top-left: Original image with foveation regions indicated; Top-right: EFIC with the upper foveation region only; Mid-left:

EFIC with the lower left foveation region only; Mid-right: EFIC with the lower right foveation region only; Bottom-left: EFIC with all the three foveation regions; Bottom-right: SPIHT uniform resolution compression.