Chapter 14 in Digital Video Image Quality and Perceptual Coding
(H. R. Wu, and K. R. Rao, eds.), Marcel Dekker Series in Signal
Processing and Communications, Nov. 2005.

Foveated Image and Video Coding

Zhou Wang and Alan C. Bovik

The human visual system (HVS) is highly space-variant in sampling, coding, processing, and understanding of visual information. The visual sensitivity is highest at the point of fixation and decreases dramatically with distance from the point of fixation. By taking advantage of this phenomenon, foveated image and video coding systems achieve increased compression efficiency by removing considerable high-frequency information redundancy from the regions away from the fixation point without significant loss of the reconstructed image or video quality.

This chapter has three major purposes. The first is to introduce the background of the foveation feature of the HVS that motivates the research effort of foveated image processing. The second is to review various foveation techniques that have been used to construct image and video coding systems. The third is to provide in more details a specific example of such systems, which delivers rate scalable codestreams ordered according to foveation-based perceptual importance, and has a wide range of potential applications such as video communications over heterogeneous, time-varying, multi-user and interactive networks.

1.1 Foveated Human Vision and Foveated Image Processing

Let us start by looking at the anatomy of the human eye. A simplified structure is illustrated in Figure 1.1. The light that passes through the optics of the eye is projected onto the retina and sampled by the photoreceptors in the retina. The retina has two major types of photoreceptors known as cones and rods. The rods

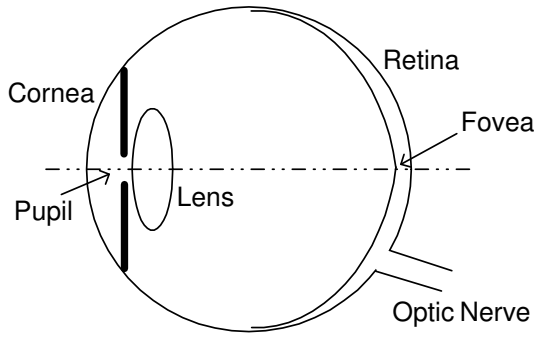


Figure 1.1: Structure of the human eye.

support achromatic vision in low level illuminations and the cone receptors are responsible for daylight vision. The cones and rods are non-uniformly distributed over the surface of the retina [1, 2]. The region of highest visual acuity is the fovea, which contains no rods but has the highest concentration of approximately 50,000 cones [2]. Figure 1.2 shows the variation of the densities of photoreceptors with retinal eccentricity, which is defined as the visual angle (in degree) between the fovea and the location of the photoreceptor. The density of the cone cells is highest at zero eccentricity (the fovea) and drops rapidly with increasing eccentricity. The photoreceptors deliver data to the plexiform layers of the retina, which provide both direct and inter-connections from the photoreceptors to the ganglion cells. The distribution of ganglion cells is also highly non-uniform as shown in Figure 1.2. The density of the ganglion cells drops even faster than the density of the cone receptors. The receptive fields of the ganglion cells also vary with eccentricity [1, 2].

The density distributions of cone receptors and ganglion cells play important roles in determining the ability of our eyes in resolving what we see. 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 information processing units in the human brain. The region around the point of fixation (or foveation point) is projected onto the fovea, sampled with the highest density, and perceived by the observer with the highest contrast sensitivity. The sampling density and the contrast sensitivity decrease dramatically with increasing eccentricity. An example is shown in Figure 1.3, where Figure 1.3(a) is the original "Goldhill" image and Figure 1.3(b) is a foveated version of that image. At certain viewing distance, if attention is focussed at the man at the lower part of the image, then the foveated and the original images are almost indistinguishable.

Despite the highly space-variant sampling and processing features of the HVS, traditional digital image processing and computer vision systems represent images on uniformly sampled rectangular lattices, which have the advantages of simple acquisition, storage, indexing and computation. Nowadays, most digital images

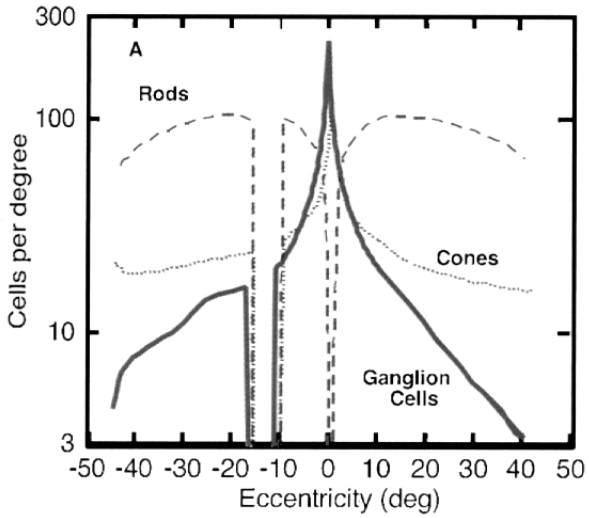


Figure 1.2: Photoreceptor and ganglion cell density versus retinal eccentricity. (From [1]).

and video sequences are stored, processed, transmitted and displayed in rectangular matrix format, in which each entry represents one sampling point. In recent years, there has been growing interest in research work on *foveated image processing* [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46], which is targeted at a number of application fields. Significant examples include image quality assessment [33, 38], image segmentation [24], stereo 3D scene perception [22], volume data visualization [9], object tracking [25], and image watermarking [42]. Nevertheless, the majority of research has been focused on foveated image and video coding, communication and related issues. The major motivation is that considerable high frequency information redundancy exists in the peripheral regions, thus more efficient image compression can be obtained by removing or reducing such information redundancy. As a result, the bandwidth required to transmit the image and video information over communication channels is significantly reduced. Foveation techniques also supply some additional benefits in visual communications. For example, in noisy communication environments, foveation provides a natural way for unequal error-protection of different spatial regions in the image and video streams being transmitted. Such an error-resilient coding scheme has shown to be more robust than protecting all the image regions equally [27, 46]. For another example, in an interactive multi-point communication environment where information about the foveated regions at the terminals of the communication networks is available, higher perceptual quality images can be achieved by applying foveated coding techniques [40].

Perfect foveation of discretely-sampled images with smoothly varying resolution

(a) (b)

Figure 1.3: Sample foveated image. (a) original "Goldhill" image; (b) foveated "Glodhill" image.

turns out to be a difficult theoretical as well as implementation problem. In the next section, we review various practical foveation techniques that approximate perfect foveation. Section 1.3 discusses a continuously rate-scalable foveated image and video coding system that has a number of good features in favor of network visual communications.

1.2 Foveation Methods

The foveation approaches proposed in the literature may be roughly classified into three categories: geometric method, filtering-based method, and multiresolution method. These methods are closely related and the third method may be viewed as a combination of the first two.

1.2.1 Geometric Methods

The general idea of the geometric methods is to make use of the foveated retinal sampling geometry. We wish to associate such a highly non-uniform sampling geometry with a spatially-adaptive coordinate transform, which we call the foveation coordinate transform. When the transform is applied to the non-uniform retinal sampling points, uniform sampling density is obtained in the new coordinate system. A typically used solution is the logmap transform [13] defined as

$$\mathbf{w} = \log(\mathbf{z} + a) \ , \tag{1.1}$$

(a) (b)

Figure 1.4: Application of foveation coordinate transform to images. (a) original image; (b) transformed image.

where a is a constant, and $\mathbf{z}$ and $\mathbf{w}$ are complex numbers representing the positions in the original coordinate and the transformed coordinate, respectively. While the logmap transform is empirical, it is shown in [34] that precise mathematical solutions of the foveation coordinate transforms may be derived directly from given retinal sampling distributions.

The foveated retinal sampling geometry can be used in different ways. The first method is to apply the foveation coordinate transform directly to a uniform resolution image, thus the underlying image space is mapped onto the new coordinate system as exemplified by Figure 1.4. In the transform domain, the image is treated as a uniform resolution image, and regular uniform-resolution image processing techniques, such as linear and non-linear filtering and compression, are applied. Finally, the inverse coordinate transform is employed to obtain a "foveatedly" processed image. The difficulty with this method is that the image pixels originally located at integer grids are moved to non-integer positions, making it difficult to index them. Interpolation and resampling procedures have to be applied in both the transform and the inverse transform domains. These procedures not only significantly complicate the system, but may also cause further distortions.

The second approach is the superpixel method [13, 16, 6, 15, 14], in which local image pixel groups are averaged and mapped into superpixels, whose sizes are determined by the retinal sampling density. Figure 1.5 shows a sophisticated superpixel look-up table given in [13], which attempts to adhere with the logmap structure. However, the number and variation of superpixel shapes make it inconvenient to manipulate. In [16], a more practical superpixel method is used, where all the superpixels have rectangular shapes. In [14], a multistage superpixel ap-

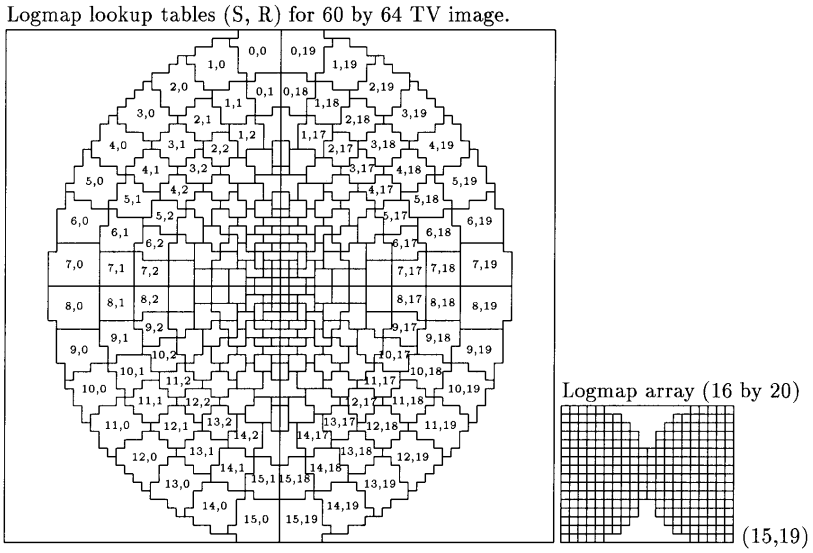


Figure 1.5: Logmap superpixel representation. The superpixel mask is applied in the pixel coordinates. (From [13])

proach is introduced, in which a progressive transmission scheme is implemented by using variable sizes of superpixels in each stage. There are two drawbacks of the superpixel methods. First, the discontinuity across superpixels is often very perceptually annoying. Blending methods are usually used to reduce the boundary effect, leading to additional computational cost. Second, when the foveation point moves, the superpixel mask has to be recalculated.

In the third method, the foveated retinal geometry is employed to guide the design of a non-uniform subsampling scheme on the uniform resolution image. An example of foveated sampling design is shown in Figure 1.6 [13]. In [23], uniform grid images are resampled with variable resolution that matches the human retina sampling density. B-Spline interpolation is then used to reconstruct the foveated images. The subsampling idea has also been used to develop foveated sensoring schemes to improve the efficiency of image and video acquisition systems [7, 13].

1.2.2 Filtering Based Methods

The sampling theorem states that the highest frequency of a signal that can be represented without aliasing is one-half of the sampling rate. As a result, the bandwidth of perceived local image signal is limited by local retinal sampling density. In the category of filtering-based foveation methods, foveation is implemented with a shift-variant low-pass filtering process over the image, where the cut-off frequency of the filter is determined by the local retinal sampling density.

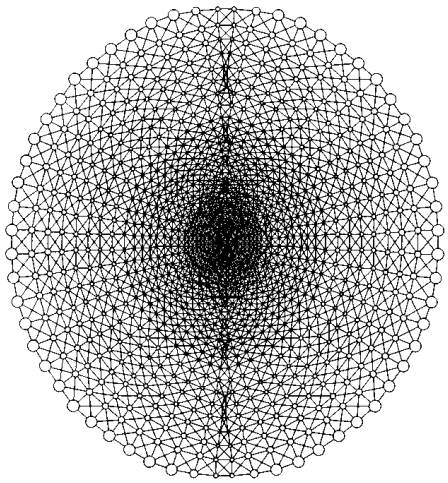


Figure 1.6: Foveated sensor distribution. (From [13])

Since retinal sampling is spatially varying (smoothly), an ideal implementation of foveation filtering would require using a different low-pass filter at each location in the image. Although such a method delivers very high quality foveated images, it is extremely expensive in terms of computational cost when the local bandwidth is low.

The filter bank method provides a flexible trade-off between the accuracy and the cost of the foveation filtering process. As illustrated in Figure 1.7 [34], a bank (finite number) of filters with varying frequency responses are first uniformly applied to the input image, resulting in a set of filtered images. The foveated image is then obtained by merging these filtered images into one image, where the merging process is space-variant according to the foveated retinal sampling density. There are a number of issues associated with the design of such filter banks and merging processes. First, the bank of filters can be either low-pass or band-pass, and thus the merging process should be adjusted accordingly. Second, there are a number of filter design problems. For instance, the filters can be designed either in the spatial or in the frequency domain. For another example, in the design of finite impulse response filters, it is important to consider the trade-offs between transition band size, ripple size, and implementation complexity (e.g., filter length). Usually, small ripple size is desired to avoid significant ringing effect. Third, since both foveation filtering and transform-based (e.g., discrete cosine transform (DCT) or wavelet-based) image compression require transforming the image signal into frequency subbands, they may be combined to reduce implementation and computational complexity. For example, only one-time DCT is used and then both foveation filtering and compression can be implemented by manipulating the DCT coefficients.

In [27, 30, 38], the filter bank method was employed as a preprocessing step

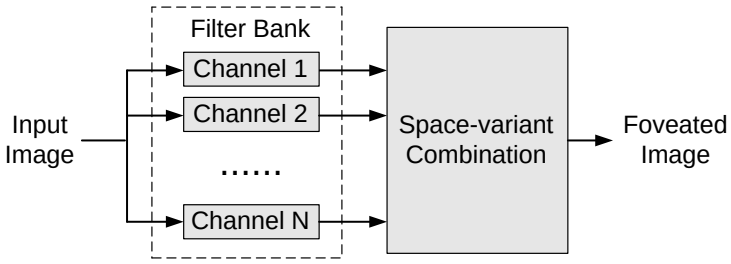


Figure 1.7: Filter bank foveation method.

before the standard video compression algorithms such as MPEG and H.26x were applied. Significant compression improvement over uniform resolution coding was obtained because a large amount of visually redundant high frequency information is removed during the foveation filtering processes. Another important advantage of this system is that it is completely compatible with the video coding standards, because no modification on the encoder/decoder of the existing video coding systems is needed, except for adding a foveation filtering unit in front of the video encoder. A demonstration of this system is available at [28].

The filter bank method was also used to build an eye tracker-driven foveated imaging system at the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin [35]. The system detects the fixation point of the subject in real-time using an eye-tracker. The detected fixation point is then used to promptly foveate the image or video being displayed on a computer monitor or projected on a large screen mounted on the wall. Since all the processes are implemented in real time, the subject feels as if he/she were watching the original image sequence instead of the foveated one, provided the calibration process has been well-performed.

In [40, 43], foveation filtering was implemented in the DCT domain and combined with the quantization processes in standard H.26x and MPEG compression. In [39], such a DCT-domain foveation method is merged into a video transcoding system, which takes compressed video streams as the input and re-encoders them into lower bit rates. Implementing these systems is indeed very challenging because the coding blocks in the current frame need to be predicted from the regions in the previous frame that may cover multiple DCT blocks and have different and varying resolution levels. It needs to point out that although the existing standard video encoders need to be modified to support foveated coding, these systems are still standard compatible in the sense that no change is necessary in order for any standard decoders to correctly decompress the received video streams.

1.2.3 Multiresolution Methods

The multiresolution method can be considered as a combination of the geometric and the filtering-based methods, in which the original uniform resolution image is transformed into different scales (where certain geometric operations such as downsampling are involved), and the image processing algorithms are applied separately at each scale (where certain filtering processes are applied).

The multiresolution method has advantages over both geometric and filtering-based methods. First, no sophisticated designs for the geometric transforms or superpixels are necessary since scaling can be implemented by simple uniform downsampling. This saves computation as well as storage space, and makes pixel indexing easy. Second, after downsampling, the number of transformed coefficients in each scale is greatly reduced. As a result, the computational cost of the filtering process decreases.

In [5], a multiresolution pyramid method [47] is applied to an uniform resolution image and a coarse-to-fine spatially adaptive scheme is then applied to select the useful information for the construction of the foveated image. In [21], a very efficient pyramid structure shown in Figure 1.8 is used to foveate images and video. In order to avoid severe discontinuities occurring across stage boundaries in the reconstructed image and video, strong blending postprocessing algorithms were employed. This system can be used for real-time foveated video coding and transmission. In [36], the system was further improved to create high quality foveated images and video that can be used to study the roles of central and peripheral vision in visual tasks such as search, navigation and reading. A demonstration of these systems and their software implementation is available at [37].

As a powerful multiresolution analysis tool, the wavelet transform has been extensively used for various image processing tasks in recent years [48]. A well-designed wavelet transform not only delivers a convenient, spatially localized representation of both frequency and orientation information of the image signal, but also allows for perfect reconstruction. These features are important for efficient image compression. In [19, 26], a non-uniform foveated weighting model in the wavelet transform domain is employed for wavelet foveation. A progressive transmission method was also suggested for foveated image communication, where the ordering of the transmitted information was determined by the foveated weighting model.

In the next section, we will mainly discuss a wavelet-based foveated scalable coding method proposed in [31, 34], where the foveated weighting model was developed by joint consideration of multiple HVS factors, including the spatial variance of the contrast sensitivity function, the spatial variance of the local visual cutoff frequency, and the variance of the human visual sensitivity in different wavelet sub-

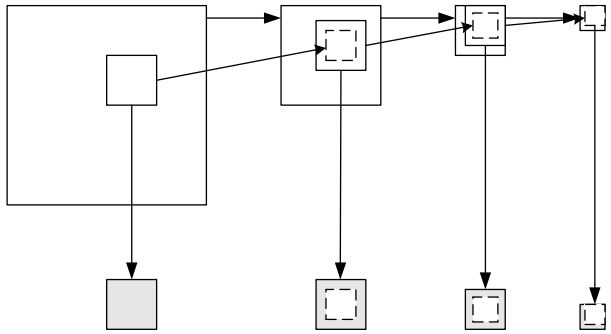


Figure 1.8: Foveated multiresolution pyramid. (Adapted from [21])

bands. The ordering of the encoded information not only depends on the foveated weighting model, but also on the magnitudes of the wavelet coefficients. This method was extended for the design of a prototype for scalable foveated video coding [44, 34]. The prototype was implemented in a specific application environment, where foveated scalable coding was combined with an automated foveation point selection scheme and an adaptive frame prediction algorithm.

1.3 Scalable Foveated Image and Video Coding

An important recent trend in visual communications is to develop continuously rate scalable coding algorithms (e.g., [49, 50, 51, 52, 53, 54, 31, 44]), which allow the extraction of coded visual information at continuously varying bit rates from a single compressed bitstream. An example is shown in Figure 1.9, where the original video sequence is encoded with a rate scalable coder and the encoded bitstream is stored frame by frame. During the transmission of the coded data on the network, we can scale, or truncate, the bitstream at any place and send the most important bits of the bitstream. Such a scalable bitstream can provide numerous versions of the compressed video at various data rates and levels of quality. This feature is especially suited for video transmission over heterogeneous, multi-user, time-varying and interactive networks such as the Internet, where variable bandwidth video streams need to be created to meet different user requirements. The traditional solutions, such as layered video (e.g., [55]), video transcoding (e.g., [56]), and simply repeated encoding, require more resources in terms of computation, storage space and/or data management. More importantly, they lack the flexibility to adapt to time-varying network conditions and user requirements, because once the compressed video stream is generated, it becomes inconvenient to change it to an arbitrary data rate. By contrast, with a continuously rate scalable codec, the data rate of the video being delivered can exactly match the available bandwidth on the network.

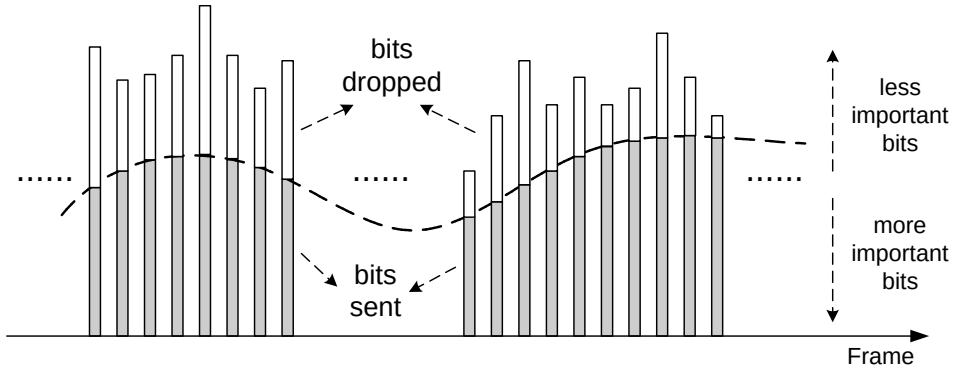


Figure 1.9: Bitstream scaling in rate scalable video communications. Each bar represents the bitstream for one frame in the video sequence. The bits in each frame are ordered according to their importance. (Adapted from [44])

The central idea of foveated scalable image and video coding is to organize the encoded bitstream to provide best decoded visual information at an arbitrary bit rate in terms of foveated perceptual quality measurement. Foveation-based HVS models play important roles in these systems. In this section, we first describe a wavelet-domain foveated perceptual weighting model, and then explain how this model is used for scalable image and video coding.

1.3.1 Foveated Perceptual Weighting Model

Psychological experiments have been conducted to measure the contrast sensitivity as a function of retinal eccentricity (e.g., [21, 57, 58]). In [21], a model that fits the experimental data was given by

$$CT(f,e) = CT_0 \exp\left(\alpha f \frac{e+e_2}{e_2}\right) , \tag{1.2}$$

where

f:	Spatial frequency (cycles/degree);
e:	Retinal eccentricity (degrees);
CT_0:	Minimal contrast threshold;
α:	Spatial frequency decay constant;
e_2:	Half-resolution eccentricity constant;
CT:	Visible contrast threshold.

The best fitting parameters given in [21] are $\alpha = 0.106$, $e_2 = 2.3$, and $CT_0 = 1/64$, respectively. The contrast sensitivity is defined as the reciprocal of the

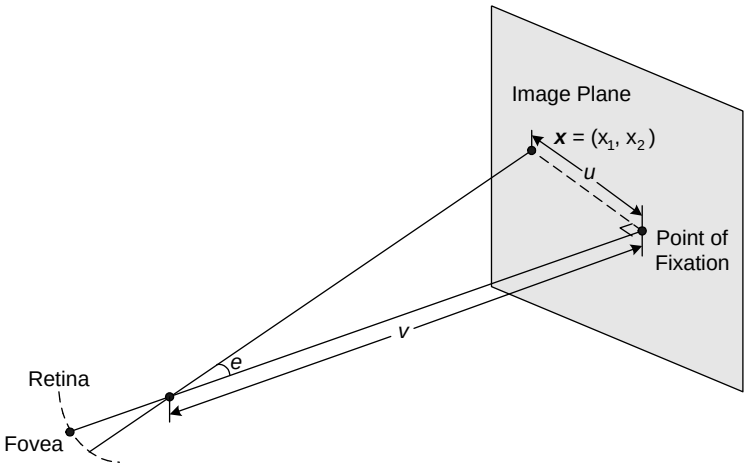


Figure 1.10: A typical viewing geometry. Here, v is the distance to the image measured in image width, and θ is eccentricity measured in degrees. (Adapted from [31])

contrast threshold: $CS(f, e) = 1/CT(f, e)$.

For a given eccentricity e, equation (1.2) 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 imperceivable. f_c can be obtained by setting CT to 1.0 (the maximum possible contrast) and solving for f:

$$f_c(e) = \frac{e_2 \ln\left(\frac{1}{CT_0}\right)}{\alpha(e + e_2)} \left(\frac{cycles}{degree}\right) . \tag{1.3}$$

To apply these models to digital images, we need to calculate the eccentricity for any given point $\mathbf{x} = (x_1, x_2)^T$ (pixels) in the image. Figure 1.10 illustrates a typical viewing geometry. For simplicity, we assume the observed image is N-pixel 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 $\mathbf{x}^f = (x_1^f, x_2^f)^T$ (pixels) and the viewing distance v (measured in image width) from the eye to the image plane are known. The distance from $\mathbf{x}$ to $\mathbf{x}^f$ is given by $d(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}^f\|_2 = [(x_1 - x_1^f)^2 + (x_2 - x_2^f)^2]^{1/2}$ (measured in pixels). The eccentricity is then calculated as

$$e(v, \mathbf{x}) = \tan^{-1}\left(\frac{d(\mathbf{x})}{Nv}\right) . \tag{1.4}$$

With (1.4), we can convert the foveated contrast sensitivity and cutoff frequency models into the image pixel domain. In Figure 1.11, we show the normalized

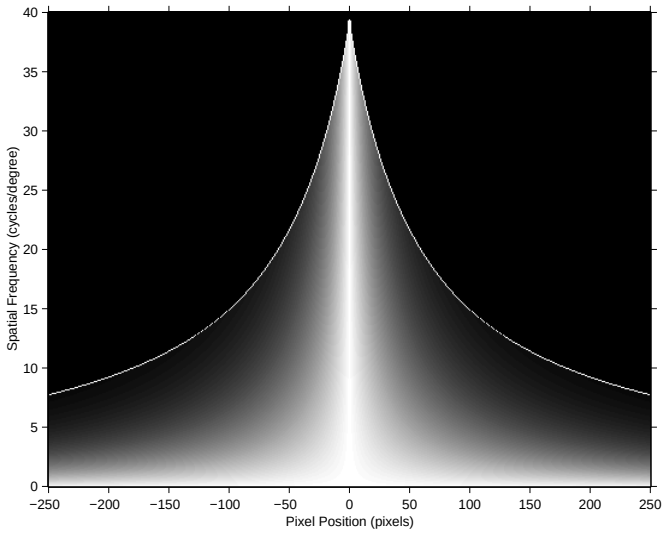


Figure 1.11: Normalized contrast sensitivity for $N = 512$ and $v = 3$. Brightness indicates the strength of contrast sensitivity and the white curves show the cutoff frequency. (Adapted from [44])

contrast sensitivity as a function of pixel position for $N = 512$ and $v = 3$. 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, which is approximately:

$$r \approx \frac{\pi N v}{180} \left(\frac{pixels}{degree} \right) \ . \tag{1.5}$$

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: $f_d(v) = r/2$. Combining this with (1.3), we obtain the cutoff frequency for a given location $\mathbf{x}$ by:

$$f_m(v, \mathbf{x}) = min(f_c(e(v, \mathbf{x})), f_d(v)) \ . \tag{1.6}$$

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

$$S_f(v, f, \mathbf{x}) = \begin{cases} \frac{CS(f, e(v, \mathbf{x}))}{CS(f, 0)} & if \quad f \leq f_m(v, \mathbf{x}) \\ 0 & otherwise \end{cases} \ . \tag{1.7}$$

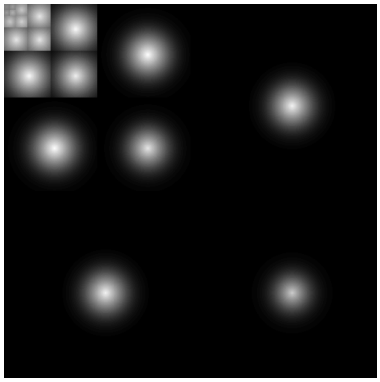


Figure 1.12: Wavelet domain importance weighting mask of a signal foveation point. Brightness (logarithmically enhanced for display purpose) indicates the importance of the wavelet coefficient. (Adapted from [44])

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

The wavelet coefficients at different subbands and locations supply information of variable perceptual importance to the HVS. In [59], 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 [59]:

$$\log Y = \log a + k(\log f - \log g_\theta f_0)^2 \tag{1.8}$$

where

Y:	Visually detectable noise threshold;
θ:	Orientation index, representing LL, LH, HH, and HL subbands, respectively;
f:	Spatial frequency (cycles/degree);
k, f_0, g_θ:	Constant parameters.

f is determined by the display resolution r and the wavelet decomposition level λ: $f = r2^{-\lambda}$. The constant parameters in (1.8) 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:

$$T_{\lambda,\theta} = \frac{Y_{\lambda,\theta}}{A_{\lambda,\theta}} = \frac{a10^{k(\log(2^\lambda f_0 g_\theta / r))^2}}{A_{\lambda,\theta}} \ , \tag{1.9}$$

where $A_{\lambda,\theta}$ is the basis function amplitude given in [59]. We define the error sensitivity in subband (λ,θ) as $S_w(\lambda,\theta) = 1/T_{\lambda,\theta}$.

For a given wavelet coefficient at position $\mathbf{x} \in \mathbf{B}_{\lambda,\theta}$, where $\mathbf{B}_{\lambda,\theta}$ 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

$$d_{\lambda,\theta}(\mathbf{x}) = 2^{\lambda} \left\| \mathbf{x} - \mathbf{x}^f_{\lambda,\theta} \right\|_2 \qquad for \quad \mathbf{x} \in \mathbf{B}_{\lambda,\theta} \ , \tag{1.10}$$

where $\mathbf{x}^f_{\lambda,\theta}$ is the corresponding foveation point in subband (λ,θ). With the equivalent distance, and also considering (1.7), we have

$$S_f(v,f,\mathbf{x}) = S_f(v, r2^{-\lambda}, d_{\lambda,\theta}(\mathbf{x})) \qquad for \quad \mathbf{x} \in \mathbf{B}_{\lambda,\theta} \ . \tag{1.11}$$

Considering both $S_w(\lambda,\theta)$ and $S_f(v,f,\mathbf{x})$, a wavelet domain foveation-based visual sensitivity model is achieved:

$$S(v,\mathbf{x}) = [S_w(\lambda,\theta)]^{\beta_1} \cdot \left[S_f(v, r2^{-\lambda}, d_{\lambda,\theta}(\mathbf{x}))\right]^{\beta_2} \quad \mathbf{x} \in \mathbf{B}_{\lambda,\theta} \ , \tag{1.12}$$

where β_1 and β_2 are parameters used to control the magnitudes of S_w and S_f, respectively.

For a given wavelet coefficient at location $\mathbf{x}$, the final weighting model is obtained by integrating $S(v,\mathbf{x})$ over v:

$$W_w(\mathbf{x}) \ = \ \int_{0^+}^{\infty} p(v) S(v,\mathbf{x}) \, dv \ , \tag{1.13}$$

where $p(v)$ is the probability density distribution of the viewing distance v [31]. Figure 1.12 shows the importance weighting mask in the DWT domain. This model can be easily generated for the case of multiple foveation points:

$$W_w(\mathbf{x}) = W^j_w(\mathbf{x}), \quad j \ \in \underset{i\in\{1,\cdots,K\}}{\arg\min} \left\{ \left\| \mathbf{x} - \mathbf{x}^f_{i,\lambda,\theta} \right\|_2 \right\} \ , \tag{1.14}$$

where K is the number of foveation points, $\mathbf{x}^f_{i,\lambda,\theta}$ is the position of the i-th foveation point in the subband (λ,θ), and $W^i_w(\mathbf{x})$ is the wavelet-domain foveated weighting model obtained with the i-th foveation point.

1.3.2 Embedded Foveation Image Coding

The embedded foveation image coding (EFIC) system [31] is shown in Figure 1.13. Firstly, the wavelet transform is applied to the original image. The foveated per-

ceptual weighting mask calculated from given foveation points or regions is then used to weight the wavelet coefficients. Next, we encode the weighted wavelet coefficients using a modified set partitioning in hierarchical trees (SPIHT) encoder, which is adapted from the SPIHT coder proposed in [51]. Finally, 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 algorithm. The foveated weighting mask is then calculated in exactly the same way as at the encoder 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 (scale) the received bitstream to obtain any bit rate image below the encoder bit rate.

The modified SPIHT algorithm employed by the EFIC system uses an embedded bit-plane coding scheme. The major purpose is to progressively select and encoder the most important remaining bit in the wavelet representation of the image. An important statistical feature of natural images that has been successfully used by the embedded zero tree wavelet (EZW) [49] and SPIHT [51] algorithms is 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. During encoding, 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. An entropy coder can be employed to further compress the output bitstream. In EFIC, the wavelet coefficients being encoded are weighted, which leads to increased dynamic range of the coefficients. This not only increases the number of scans, but also increases the number of bits to encode the large coefficients. The modified SPIHT algorithm employed by EFIC limits the maximum number of bits for each coefficient and scans only the strongly weighted coefficients in the first several scans. Both of these modifications reduce computational complexity and increase the overall coding efficiency.

Figure 1.14 shows the 8 bits/pixel gray scale “Zelda” image encoded with SPIHT and EFIC, where the foveated region is at the center of the image. At a low bit-rate of 0.015625 bits/pixel with compression ratio (CR) equaling 512:1, the mouth, nose, and eye regions are hardly recognizable in the SPIHT coded

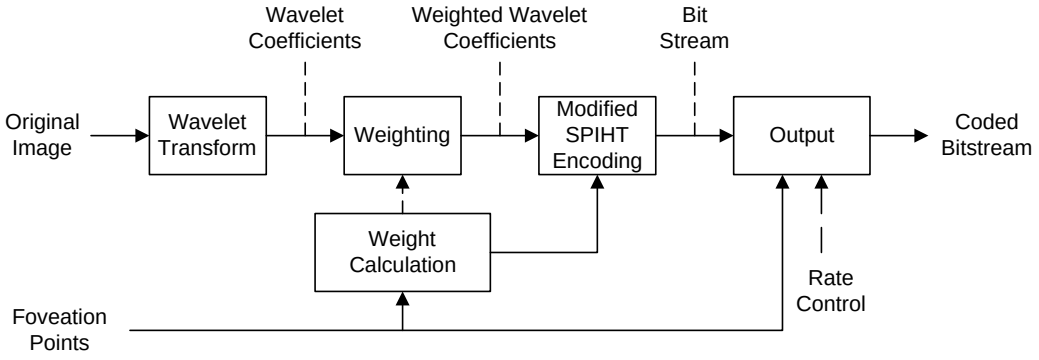


Figure 1.13: EFIC encoding system. (Adapted from [31])

image, whereas those regions in the EFIC coded image exhibit some useful information. At a medium bit-rate of 0.0625 bits/pixe (CR = 128:1), SPIHT still decodes a quite blurry image, while EFIC gives much more detailed information over the face region. Increasing the bit-rate to as high as 0.25 bits/pixel (CR = 25), the EFIC coded image approaches uniform resolution. The decoded SPIHT and EFIC images both have high quality and are almost indistinguishable. More demonstration images for EFIC can be found at [32].

The EFIC decoding procedure can also be viewed as a progressive foveation filtering process with gradually decreasing foveation depth. The reason may be explained as follows: Note that the spectra of natural image signals statistically follow the power law $1/f^p$ (see [60] for a review). As a result, the low-frequency wavelet coefficients are usually larger than the high-frequency ones, thus generally have better chances to be reached earlier in the embedded bit-plane coding process. Also notice that the foveated weighting process shifts down the bit-plane levels of all the coefficients in the peripheral regions. Therefore, at the same frequency level, the coefficients at the peripheral regions generally occupy lower bit-planes than the coefficients at the region of fixation. If the available bit-rate is limited, then the embedded bit-plane decoding process corresponds to applying a higher-bandwidth low-pass filter to the region of fixation and a lower-bandwidth low-pass filter to the peripheral regions, thereby foveates the image. With the increase of bit-rate, more bits for the high-frequency coefficients in the peripheral regions are received, thus the decoded image becomes less foveated. This is well demonstrated by the EFIC coded images shown in Figure 1.14.

1.3.3 Foveation Scalable Video Coding

The foveated scalable video coding (FSVC) system [44] follows the general method of motion estimation/motion compensation-based video coding. It first divides the

Figure 1.14: "Zelda" image compressed with SPIHT and EFIC algorithms. (a) SPIHT compressed image, compression ratio (CR) = 512:1; (b) EFIC compressed image, CR = 512:1; (c) SPIHT compressed image, CR = 128:1; (b) EFIC compressed image, CR = 128:1; (e) SPIHT compressed image, CR = 32:1; (f) EFIC compressed image, CR = 32:1. (Adapted from [31])

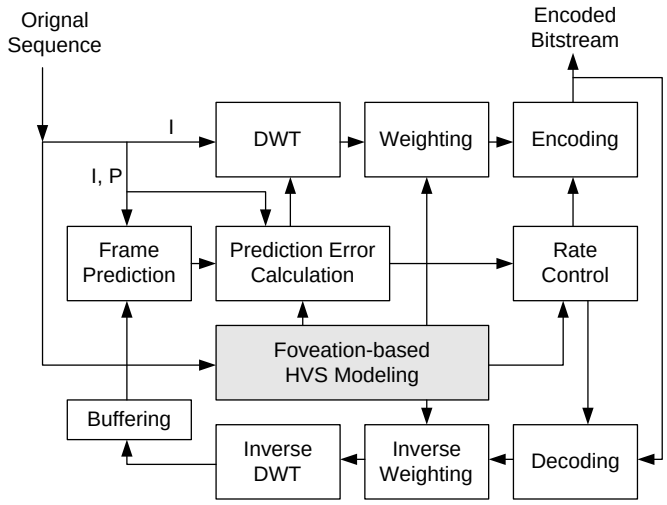


Figure 1.15: FSVC encoding system. (Adapted from [44])

input video sequence into groups of pictures (GOPs). Each GOP has one intra-coding frame (I frame) at the beginning and the rest are predictive coding frames (P frames). The diagram of the encoding system is shown in Figure 1.15. The I frames are encoded using the same way as in the EFIC algorithm described above. The encoding of P frames is more complicated and is different from other video coding algorithms in that it uses two instead of one version of the previous frames. One is the original previous frame and the other is a feedback decoded version of the previous frame. The final prediction frame is the weighted combination of the two motion compensated prediction frames. The combination is based on the foveated weighting model.

The prototype FSVC system allows to select multiple foveation points, mainly to facilitate the requirements of large foveation regions and multiple foveated regions of interest. It also reduces the search space of the foveation points by dividing the image space into blocks and limiting the candidate foveation points to the centers of blocks. This strategy not only decreases implementation and computational complexity, but also reduces the number of bits needed to encode the positions of the foveation points. In practice, the best way of foveation point(s) selection is application dependant. The FSVC prototype is very flexible such that different foveation point selection schemes can be applied to a single framework.

We implemented the FSVC prototype in a specific application environment for video sequences with human faces. A face-foveated video coding algorithm is useful to effectively enhance the visual quality in specific video communication environments such as videoconferencing.

The methods to choose foveation points for I frames and P frames are different. In the I frames, a face detection algorithm similar to that in [61] is used, which

detects possible face regions by the skin color information [62] and uses a binary template matching method to detect human faces in the skin-color regions. A different strategy is used for P frames, where we concentrate on the regions in the current P frame that provide us with new information from its previous frame, in which the prediction errors are usually larger than other regions. The potential problem of this method is that the face regions may lose fixation. To solve this problem, an unequal error thresholding method is used to determine foveation regions in P frames, where a much smaller prediction error threshold value is used to capture the changes occurring in the face regions. In Figure 1.16, we show five consecutive frames in the "Silence" sequence and the corresponding selected foveation points, in which the first frame is an I frame and the rest are P frames.

In fixed-rate motion compensation-based video coding algorithms, a common choice is to use the feedback decoded previous frame as the reference frame for the prediction of the current frame. This choice is infeasible for continuously scalable coding because the decoding bit rate may be different from the encoding bit rate and is unavailable to the encoder. In [52], a low base rate is defined and the decoded and motion compensated frame at the base rate is used as the prediction. This solution avoids the significant error propagation problems, but when the decoding bit rate is much higher than the base rate, large prediction errors may occur and the overall coding efficiency may be seriously affected. A new solution to this problem is used in the FSVC system, where the original motion compensated frame and the base rate decoded and motion compensated frame are adaptively combined using the foveated weighting model. The idea is to assign more weight to the base rate motion compensated frame for difficult prediction regions, and more weight to the original motion compensated frame for easy prediction regions. By using this method, error propagation becomes a small problem, while at the same time, better frame prediction is achieved, which leads to smaller prediction errors and better compression performance.

Figure 1.16 shows the FSVC compression results of the "Silence" sequence. It can be observed that the spatial quality variance in the decoded image sequences is well adapted to the time-varying foveation point selection scheme. Figure 1.17 demonstrates the scalable feature of the FSVC system, which shows the reconstructed 32nd frame of the "Salesman" video sequence decoded at 200, 400 and 800 Kbits/sec, respectively. The reconstructed video sequences are created from the same FSVC-encoded bitstream by truncating the bitstream at different places. Similar to Figure 1.14, the decoded images exhibit decreased foveation depth with increasing bit rate.

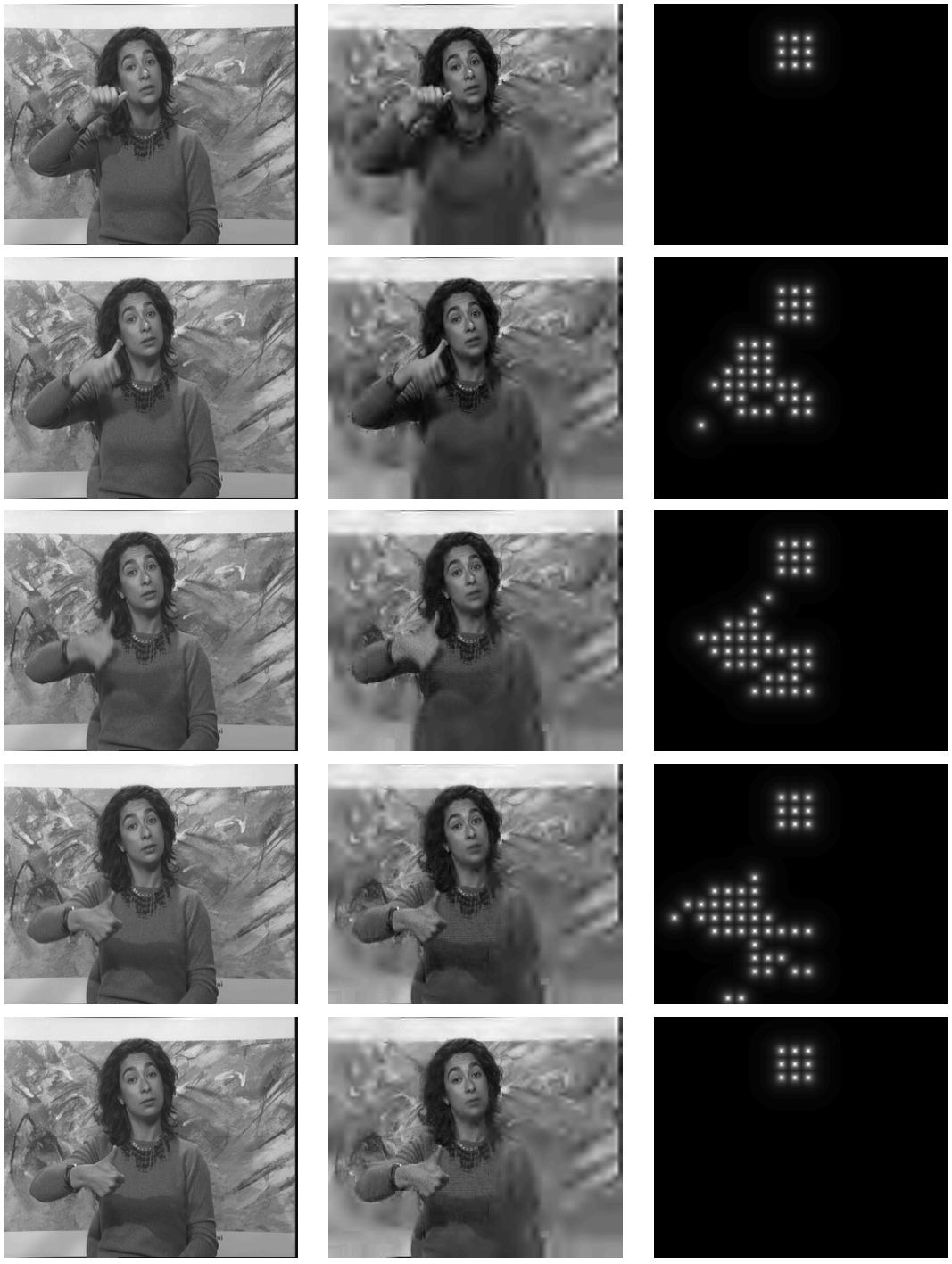
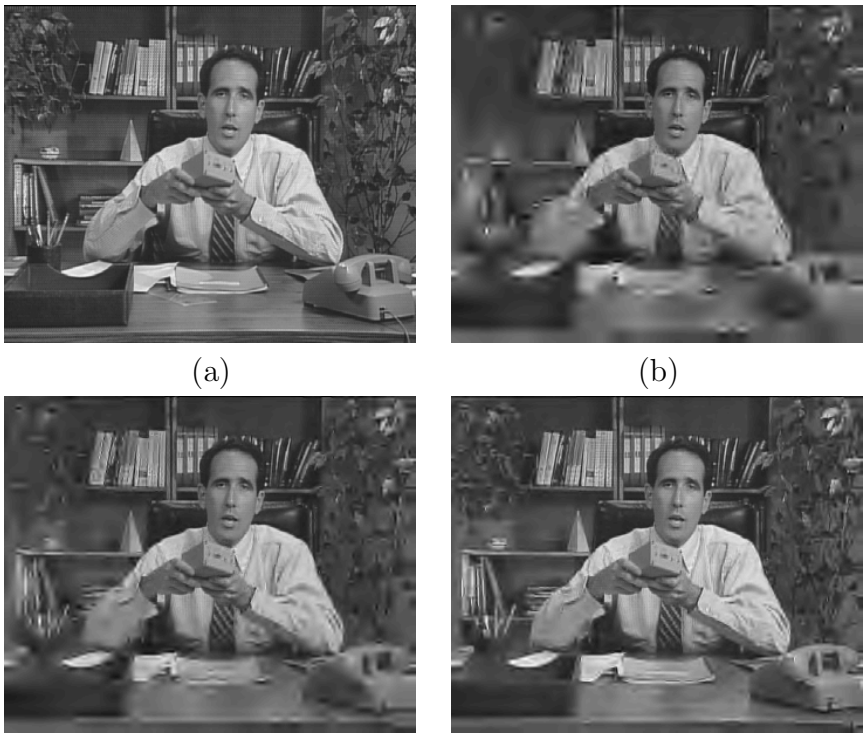


Figure 1.16: Consecutive frames of the "Silence" sequence (left); the FSVC compression results at 200 Kbits/sec (middle); and the selected foveation points (right). (Adapted from [44])



(a) (b)

(c) (d)

Figure 1.17: Frame 32 of the "Salesman" sequence (a) compressed using FSVC at 200 Kbits/sec (b), 400 Kbits/sec (c), and 800 Kbits/sec (d), respectively. (Adapted from [44])

1.4 Discussions

This chapter first introduces the background and motivations of foveated image processing, and then reviews the various foveation techniques that are used for the development of image and video coding systems. To give examples on specific implementations of such systems, we described in more details the EFIC and the FSVC systems, which supply continuously rate-scalable codestrems ordered according to foveation-based perceptual importance. Such systems have a number of potential applications.

One direct application is network image browsing. There are two significant examples. In the first example, prior to using the encoding algorithm, the foveation point(s) are predetermined. The coding system then encodes the image with high bit-rate and high quality. One copy of the encoded bitstream is stored at the server side. When the image is required by a client, the server sends the bitstream to the client progressively. The client can stop the transmission at any time once the reconstructed image quality is satisfactory. In the second example, the foveation point(s) are unknown to the server before transmission. Instead of a fully encoded bitstream, a uniform resolution coarse quality version of the image is precomputed and stored at the server side. The client first sees the coarse version of the image and clicks on the point of interest in that image. The selected point of interest is sent back to the server and activates the scalable foveated encoding algorithm. The encoded bitstream that has a foveation emphasis on the selected point of interest is then transmitted progressively to the client.

Another application is network videoconferencing. Compared with traditional videoconferencing systems, a foveated system can deliver lower data rate video streams since much of the high frequency information redundancy can be removed in the foveated encoding process. Interactive information such as the locations of the mouse, touch screen and eye-tracker can be sent as feedback information to the other side of the network and used to define the foveation points. Face detection and tracking algorithm may also help to find and adjust the foveation points. Furthermore, in a highly heterogeneous network, the available bandwidth can change dramatically between two end users. A fixed bit-rate video stream would either be terminated suddenly (when the available bandwidth drops below the fixed encoding bit-rate) or suffer from the inefficient use of the bandwidth (when the fixed bit-rate is lower than the available bandwidth). By contrast, a rate scalable foveated videoconferencing system can deal with these problems more smoothly and efficiently.

The most commonly used methods for robust visual communications on noisy channels are error resilience coding at the source or channel coders and error concealment processing at the decoders [63]. Scalable foveated image and video stream

provides us with the opportunity to do a better job by taking advantage of its optimized ordering of visual information in terms of perceptual importance. It has been shown that significant improvement can be achieved by unequal error protection for scalable foveated image coding and communications [41].

Active networks are a hot research topic in recent years [64]. It allows the customers to send not only static data but also programs that are executable at the routers or switches within the network. An active network becomes more useful and effective for visual communications if an intelligent scheme is employed to modify the visual contents being delivered in a smart and efficient way. The properties of scalable foveated image/video streams provide a good match to the features of active networks because the bit rate of the video stream can be adjusted according to the network conditions monitored at certain routers/switches inside the network (instead of at the sender side), and the feedback foveation information (points and depth) at the receiver side may also be dealt with at the routers/switches. This may result in quicker responses that benefit real-time communications.

Finally, a common critical issue in all foveated image processing applications is how the foveation points or regions should be determined. Depending on the application, this may be done either interactively or automatically. In the interactive method, an eye tracker is usually used to track the eye movement and send the information back to the foveated imaging system in real time. In most application environments, however, the eye tracker is not available or is inconvenient. A more practical way is to ask the users to indicate fixation points using a mouse or touch screen. Another 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. Automatical determination of foveation points is itself a difficult but interesting research topic, and is closely related to psychological visual search research (see [65] for a review). In the image processing literature, there also has been previous research towards understanding high level and low level processes in deciding human fixation points automatically (e.g., [22, 66, 44, 67]). High level processes are usually context dependent and involve a cognitive understanding of the image and video being observed. For example, once a human face is recognized in an image, the face area is likely to become a heavily fixated region. In a multimedia environment, audio signals may also be linked to image objects in the scene and help to determine foveation points [34]. Low level processes determine the points of interest using simple local features of the image [66, 67]. In [22], three-dimensional depth information is also employed to help find foveation points in an active stereo vision system. Although it is argued that it is always difficult to decide foveation points automatically, we believe that it is feasible to establish a statistical model that predicts them in a measurably effective way.

Bibliography

[1] W. S. Geisler and M. S. Banks, "Visual performance," in *Handbook of Optics*, M. Bass, ed., McGraw-Hill, 1995.

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

[3] E. L. Schwartz, "Spatial mapping in primate sensory projection: Analytic structure and relevance to perception," *Biological Cybernetics*, vol. 25, pp. 181-194, 1977.

[4] E. L. Schwartz, "Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding," *Vision Research*, vol. 20, pp. 645-669, 1980.

[5] P. J. Burt, "Smart sensing within a pyramid vision machine," *Proc. IEEE*, vol. 76, pp. 1006-1015, Aug. 1988.

[6] C. Bandera and P. Scott, "Foveal machine vision systems," *IEEE Inter. Conf. System, Man and Cybernetics*, pp. 596-599, Nov. 1989.

[7] A. S. Rojer and E. L. Schwartz, "Design considerations for a space-variant visual sensor with complex-logarithmic geometry," *IEEE Inter. Conf. Pattern Recognition*, vol. 2, pp. 278-285, 1990.

[8] C. Weiman, "Video compression via a log polar mapping," *Real Time Image Processing II*, Proc. SPIE, vol. 1295, pp. 266–277, 1990.

[9] M. Levoy and R. Whitaker, "Gaze-Directed Volume Rendering," *Computer Graphics*, vol. 24, no. 2, pp. 217-223, 1990.

[10] Y. Y. Zeevi and E. Shlomot, "Nonuniform sampling and antialiasing in image representation," *IEEE Trans. Signal Processing*, vol. 41, no. 3, pp. 1223-1236, Mar. 1993.

[11] P. L. Silsbee, A. C. Bovik and D. Chen, "Visual pattern image sequence coding," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 3, no. 4, pp. 291-301, Aug. 1993.

[12] A. Basu, A. Sullivan and K. J. Wiebe, "Variable resolution teleconferencing," *IEEE Inter. Conf. Systems, Man, and Cybernetics*, pp. 170-175, Oct. 1993.

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

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

[15] P. Camacho, F. Arrebola and F. Sandoval, "Shifted fovea multiresolution geometries," *IEEE Inter. Conf. Image Processing*, vol. 1, pp. 307-310, 1996.

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

[17] T. L. Arnow and W. S. Geisler, "Visual detection following retinal damage: Prediction of an inhomogeneous retino-cortical model," *Human Vision and Electronic Imaging*, Proc. SPIE, vol. 2674, pp. 119-130, 1996.

[18] T. H. Reeves and J. A. Robinson, "Adaptive foveation of MPEG video," *ACM Multimedia*, pp. 231-241, 1996.

[19] E.-C. Chang and C. Yap, "A wavelet approach to foveating images," *ACM Symposium on Computational Geometry*, pp. 397-399, June 1997.

[20] A. Basu and K. J. Wiebe, "Videoconferencing using spatially varying sensing with multiple and moving fovea," *IEEE Trans. Systems, Man and Cybernetics*, vol. 28, no. 2, pp. 137-148, Mar. 1998.

[21] W. S. Geisler and J. S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication," *Human Vision and Electronic Imaging*, Proc. SPIE, vol. 3299, pp. 294-305, July 1998.

[22] W. N. Klarquist and A. C. Bovik, "FOVEA: A foveated vergent active stereo vision system for dynamic three-dimensional scene recovery," *IEEE Trans. Robotics and Automation*, vol. 14, no. 5, pp. 755-770, Oct. 1998.

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

[24] J. M. Kinser, "Foveation from pulse images," *IEEE Inter. Conf. Information Intelligence and Systems*, pp. 86-89, 1999.

[25] R. Etienne-Cummings, J. van der Spiegel, P. Mueller and M.-Z. Zhang, "A foveated silicon retina for two-dimensional tracking," *IEEE Trans. Circuits and Systems II*, vol. 47, no. 6, pp. 504-517, June 2000.

[26] E.-C. Chang, S. Mallat and C. Yap, "Wavelet foveation," *Journal of Applied and Computational Harmonic Analysis*, vol. 9, no. 3, pp. 312-335, Oct. 2000.

[27] S. Lee, *Foveated video compression and visual communications over wireless and wireline networks*, Ph.D. dissertation, Dept. of ECE, University of Texas at Austin, May 2000.

[28] S. Lee and A. C. Bovik, "Foveated video demonstration," Dept. of ECE, University of Texas at Austin, *http://live.ece.utexas.edu/research/foveated_video/demo.html*, 2000.

[29] S. Daly, K. Matthews and J. Ribas-Corbera, "As plain as the noise on your face: Adaptive video compression using face detection and visual eccentricity models," *Journal of Electronic Imaging*, vol. 10, pp. 30-46, Jan. 2001.

[30] S. Lee, M. S. Pattichis and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Trans. Image Processing*, vol. 10, no. 7, pp. 977-992, July 2001.

[31] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Trans. Image Processing*, vol. 10, no. 10, pp. 1397-1410, Oct. 2001.

[32] Z. Wang and A. C. Bovik, "Demo images for 'embedded foveation image coding'," *http://www.cns.nyu.edu/~zwang/files/research/efic/demo.html*, 2001.

[33] Z. Wang, A. C. Bovik, and L. Lu, "Wavelet-based foveated image quality measurement for region of interest image coding," *IEEE Inter. Conf. Image Processing*, vol. 2, pp. 89-92, Oct. 2001.

[34] Z. Wang, *Rate scalable foveated image and video communications*, Ph.D. dissertation, Dept. of ECE, University of Texas at Austin, Dec. 2001.

[35] U. Rajashekar, Z. Wang and A. C. Bovik, "Real-time foveation: An eye tracker-driven imaging system," *http://live.ece.utexas.edu/research/realtime_foveation/*, 2001.

[36] J. S. Perry and W. S. Geisler, "Gaze-contingent real-time simulation of arbitrary visual fields," *Human Vision and Electronic Imaging*, B. Rogowitz and T. Pappas, eds., Proc. SPIE, vol. 4662, San Jose, CA, 2002.

[37] "Space variant imaging," Center for Perceptual Systems, University of Texas at Austin, *http://www.svi.cps.utexas.edu/*, 2002.

[38] S. Lee, M. S. Pattichis and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 129-132, Mar. 2002.

[39] S. Liu, *DCT domain video foveation and transcoding for heterogeneous video communication*, Ph.D. dissertation, Dept. of ECE, University of Texas at Austin, May 2002.

[40] H. R. Sheikh, S. Liu, Z. Wang and A. C. Bovik, "Foveated multipoint videoconferencing at low bit rates," *IEEE Inter. Conf. Acoust., Speech, and Signal Processing*, vol. 2, pp. 2069-2072, May 2002.

[41] M. F. Sabir, *Unequal error protection for scalable foveated image communication*, Master's thesis, Dept. of ECE, University of Texas at Austin, May 2002.

[42] A. Koz and A. Alatan, "Foveated image watermarking," *IEEE Inter. Conf. Image Processing*, vol. 3, pp. 661-664, Sept. 2002.

[43] H. R. Sheikh, B. L. Evans and A. C. Bovik, "Real-time foveation techniques for low bit rate video coding," *Real-time Imaging*, vol. 9, no. 1, pp. 27-40, Feb. 2003.

[44] Z. Wang, L. Lu and A. C. Bovik, "Foveation scalable vidoe coding with automatic fixation selection," *IEEE Trans. Image Processing*, vol. 11, no. 2, pp. 243-254, Feb. 2003.

[45] S. Lee and A. C. Bovik, "Fast algorithms for foveated video processing," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 13, no. 2, pp. 149 -162, Feb. 2003.

[46] S. Lee, C. Podilchuk, V. Krishnan and A. C. Bovik, "Foveation-based error resilience and unequal error protection over mobile networks," *Journal of VLSI Signal Processing*, vol. 34, no. 1/2, pp. 149-166, May 2003.

[47] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Communications*, vol. 31, pp. 532-540, Apr. 1983.

[48] S. G. Mallat, *A wavelet tour of signal processing*, Academic Press, 2nd edition, Sep. 1999.

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

[50] D. Taubman and A. Zakhor, "Multirate 3-D subband coding of video," *IEEE Trans. Image Processing*, vol. 3, pp. 572-588, Sep. 1994.

[51] 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, Jun 1996.

[52] K. S. Shen and E. J. Delp, "Wavelet based rate scalable video compression," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 109-122, Feb. 1999.

[53] S.-J. Choi and J. W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. Image Processing*, vol. 8, no. 2, pp. 155-167, Feb. 1999.

[54] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards, and Practice*, Kluwer Academic Publishers, Nov. 2001.

[55] S. Aramvith and M.-T. Sun, "MPEG-1 and MPEG-2 video standards," in *Handbook of Image and Video Processing*, A. Bovik, ed., Academic Press, May 2000.

[56] H. Sun, W. Kwok and J. W. Zdepski, "Architectures for MPEG compressed bitstream scaling," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 6, no. 2, pp. 191-199, Apr. 1996.

[57] 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.

[58] 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.

[59] 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.

[60] D. L. Ruderman, "The statistics of natural images," *Network: Computation in Neural Systems*, vol. 5, pp. 517-548, 1996.

[61] H. Wang and S.-F. Chang, "A highly efficient system for automatic face region detection in MPEG video," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 7, no. 4, pp. 615-628, Aug. 1997.

[62] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Trans. Multimedia*, vol. 1, no. 3, pp. 264-277, Sep. 1999.

[63] J. D. Villasenor, Y.-Q. Zhang and J. Wen, "Robust video coding algorithms and systems," *Proc. IEEE*, vol. 87, pp. 1724-1733, Oct. 1999.

[64] D. L. Tennenhouse, J. M. Smith, W. D. Sincoskie, D. J. Wetherall and G. J. Mindenk, "A survey of active network research," *IEEE Communications Magazine*, vol. 35, Jan. 1997.

[65] L. Itti, *Models of bottom-up and top-down visual attention*, Ph.D. dissertation, California Institute of Technology, 2000.

[66] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, pp. 970-982, Sep. 2000.

[67] U. Rajashekar, L. K. Cormack and A. C. Bovik, "Image features that draw fixations," *IEEE Inter. Conf. Image Processing*, vol. 3, pp. 313-316, Sep. 2003.