Fast algorithms for foveated video processing

Sanghoon Lee[1] and Alan C. Bovik[2], *Fellow, IEEE*

Abstract

This paper explores the problem of communicating high-quality, foveated video streams in real time. Foveated video exploits the nonuniform resolution of the human visual system by preferentially allocating bits according to proximity to assumed visual fixation points, thus delivering perceptually high quality at greatly reduced bandwidths. Foveated video streams possess specific data density properties that can be exploited to enhance the efficiency of subsequent video processing. Here we exploit these properties to construct several efficient foveated video processing algorithms: foveation filtering(local bandwidth reduction), motion estimation, motion compensation, video rate control and video postprocessing. Our approach leads to enhanced computational efficiency by interpreting nonuniform-density foveated images on the uniform domain and by using a foveation protocol between the encoder and the decoder.

Keywords: foveated video, foveated video processing, foveated video algorithms

[1] Yonsei University, Seoul, Korea. E-mail : sanghoon90@yahoo.com [2]The Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin.

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

I. Introduction

When the human eye focuses at a point on a surface, or on an image of a surface (called the "fixation point"), a variable-resolution image is transmitted into the brain via the nonuniform sampling of the visual signal by the retinal photoreceptors. If the fixation point of the observer can be ascertained, then mathematical models of this *foveation property* of the eye can be exploited to enhance image and video processing algorithms. For example, high spatial frequencies that occur locally and away from the fixation point can be removed or attenuated to increase the compressibility of the image or video data. An image or video stream that has been processed in this way is called *foveated*. If the fixation point is accurately determined, and the reduction in local high spatial frequencies is accomplished according to the foveation law of the eye and accounting for the viewing geometry, then the image or video stream can be made to be perceptually equivalent to the original. Fig. 1 (a) and (b) show an original and a foveated version of an image, where the foveation point (the presumed fixation point) is indicated by an "x". If the viewer fixates at this point from an appropriate viewing distance, the perceived image (b) will differ little from the original image (a) (some difference will likely be perceived by the reader, since the journal reproduction of the image is small).

The increased compressibility of foveated image and video data can be understood in terms of entropy. Given a foveated image, a coordinate transformation can be found such that maps it into a uniform-resolution image. Fig. 1 (d) shows the foveated image (b) remapped into uniformly sampled coordinates, while Fig. 1 (c) shows the local bandwidth of Fig. 1 (b). We define the attainable compression gain using foveation to be the entropy that is saved relative to the original image entropy. Let $H(X)$ be the average entropy of the pixels over the uniform domain in Fig. 1 (a) and (d), and let A_o and A_c be the area of Fig. 1 (a) and (d) respectively. Then, the total saved entropy can be expressed as a function of the area difference $(A_o - A_c)$ $H(X)$ where $A_o \geq A_c$.

In [1], a new measure for assessing the quality of foveated video streams was introduced: the *foveal signal-to-noise ratio* (FSNR) objectively evaluates video quality under the assumption that the video is being viewed by a foveated visual systems such as the eye. In [2][3], an optimal rate control algorithm was introduced for maximizing the FSNR of foveated and compressed video streams. Foveation has been shown to be an effective means for deeply scaling down the range of video transmission rates that yield efficient channel adaptation, high visual performance

and low transmission delays [4]. In [5], a foveation-based error resilience scheme and an associated method for unequal error protection were introduced.

Here we introduce several efficient foveated video processing algorithms that profit from the nonuniform resolution of the foveated data. Several stages of processing in the video communications pathway are studied: foveation filtering (space-variant filtering to create foveated video), motion estimation, motion compensation, rate control and postprocessing/enhancement. These tasks are approached in a aggregate manner by interpreting the foveated video over the uniform domain. Optimal or near-optimal performance is attained by invoking the quality metric FSNR, Parseval's theorem, the Whittaker-Shannon sampling theorem and a foveation protocol.

II. Overview of Foveated Image/Video

A. Quality Assessment

In order to coincide with the sampling characteristics of the human vision [6][7][8], picture quality should be assessed based on the distortion between the original image and the reconstructed image formed by the nonuniform sampling of the retinal neurons, instead of the image displayed on the monitor. Several psychophysical studies have been aimed in this direction, using simple image presentations (e.g. sinusoidal gratings) [9][10], and expressing image quality in terms of the display parameters: resolution, contrast, luminance, display size, viewing distance and retinal eccentricity [11][12]. In order to capture the spatially-varying response of the human visual system (HVS), a local bandlimited contrast has been introduced [13][14]. In [15][14], the contrast sensitivity function (CSF) was used as a weighting function for noise measurement and the error measurement criterion is the WSNR (weighted SNR).

However, those quality assess metrics were designed to quantify visual quality by analyzing the input image with huge computation complexity. Therefore, it is difficult to be embedded in real-time image/video compression algorithms as a quality processing criterion. In terms of simplicity and feasibility in real-time video implementation, the most commonly-applied objective error measurements have been the mean square error (MSE) and the mean absolute error (MAE). However, both the MSE and (to a slightly lesser degree) the MAE correlate poorly with subjective quality measurements.

For point-to-point visual communications, an effective quality metric should be attained by taking into account only the spatial resolution variation of the HVS. In [16], a foveated quality

metric called VRMAE (variable resolution MAE) has been developed. In the metric, a weighting factor for each pixel was obtained according to the distance from the closest foveation point. In the VRMAE, the original image was used as the reference image without introducing the concept of the foveated image. In this case, large amount MAEs may occur at peripheral areas between the original image and the reconstructed image, which is already distorted by taking the VR (variable resolution) transform. Thus, it should be difficult to quantify the localized visual quality accurately.

In [1][2], a quality metric $FSNR$ was defined, where a foveated image was obtained based on the local bandwidth of the HVS. The FSNR was calculated between the foveated image and its reconstructed image, or the original image and the reconstructed image when foveation filtering was not used. Also, the performance of the FSNR was successfully demonstrated as an objective quality metric. One more important property of the FSNR is its computational simplicity. Thus, it can be embedded in real-time video processing for optimization on rate control and motion estimation algorithms. Since the FSNR is the quality metric measured over virtual coordinates associated with the nonuniform property of the HVS, it can be used as a reference needed to remove the computational redundancy over cartesian coordinates without reducing the image size by taking any coordinate transform.

B. Foveation Filtering

B.1 Conformal Mapping and VR Transform

In [17][18], it was demonstrated that the anatomical structure of the striate cortex is simply characterized by the complex logarithmic mapping. This transformation is described as a conformal mapping of points on the polar(retinal) plane (r=$\sqrt{x_1^2 + x_2^2}$, $\theta = tan^{-1} x_2/x_1$) onto a cartesian (cortical plane) ($\eta = \log r$, $\gamma = \theta$). In [19], this topology was simulated by means of a discrete distribution of elements whose sampling distance increases linearly with eccentricity from the fovea. In vision systems, the foveation process has been utilized to reduce visual data by conformal mapping [20][21][22]. Unfortunately, reconstructing an image from its logpixels results in sharp boundaries between the regions corresponding to adjacent logpixels. The sharp boundaries manifest themselves as degradations of the original image.

In [23][24], a compressed image was generated using the VR transform, and JPEG compression was additionally applied to the compressed image over the transform domain. Then, the

decompressed image was obtained after the inverse transform. Since the reduced version of the image was used for the JPEG compression, the computational overhead was effectively reduced. However, when the local sampling rate during the mapping is less than the Nyquist sampling rate in cartesian coordinates, aliasing effects can occur and additional noise can be added in the process of the VR / inverse VR transform.

B.2 Bandlimited Signal Representation

In [25], it was shown that the human eye can experience the phenomenon of aliasing when the spatial frequency of pattern formed over the retina is higher than the cone Nyquist frequency. In [26], a sampling rate is decided at each hexagonal unit according to the eccentricity (visual angle). Therefore, a foveated image may be interpreted as having position-varying local bandwidths.

There has been some research directed towards analyzing this spatially-variable information content[27][28]. In [27], a nonuniform sampling theorem was derived by expressing the uniformity as a coordinate transformation applied to a uniform tessellation. In [28], a space of locally bandlimited signals was defined and used to process pyramidal images, where the image sampling density monotonically decreased as a function of radial distance from an area-of-interest. One approach to analyzing such spatio-spectral signals is via the AM-FM transform [29], which generalizes the spatio-spectral representation across curvilinear coordinate mappings. However, due to the computation complexity, it is difficult to apply those algorithms in real-time video processing.

B.3 Position-Frequency Representation

Since the goal of foveation processing is to selectively remove undetectable high frequency components according to position, both position and frequency information are required. Widely-used joint position-frequency representations are the short-time Fourier transform (STFT) and the wavelet transform (WT)[30]. It is well-known that Gabor filters are optimal in the sense that they minimize the time-frequency uncertainty product, and thus provide excellent time-frequency localization[31]. In [32], a signal is nonuniformly down-sampled, and a foveated signal is reconstructed using a Gabor expansion. However, this Gabor scheme is also a logpixel technique with the boundary problems that we have just described.

The discrete WT (DWT) relies on the iterative use of a fixed low-pass filter. This means that the low-pass filter approximation for the first transform level is much worse than the approxima-

tion for the second transform level, and the approximation keeps degrading with the transform level [33]. Thus, using the DWT, we cannot approximate the response of an arbitrary low-pass filter. Foveation filtering using the WT has been demonstrated in [34]. In this approach, weighted translation is used to implement foveation filtering ; however, the foveated image exhibited severe blocking artifacts. In another approach for implementing DWT-based foveation filtering, the bitstream ordering procedure is based on a visual important weighted model over the DWT domain [35]. However, since the WT is embedded in a particular image processing algorithm, it is difficult to adapt to DCT-based standard video.

C. Motivation of this paper

In this paper, we implement foveation filtering having low pass filters with continuously varying cutoff frequencies. For discrete images, we are forced to use a fixed set of cutoff frequencies, yet, unlike the WT or the STFT, we allow for an arbitrary cutoff frequency. Using such low pass filters, the performance of the algorithm does not depend on the cutoff frequency (as in the WT). As long as the local bandwidth transition is monotonically changed, position-varying lowpass filtering can be utilized to compute foveated images that better approximate the human visual system. Another merit of this approach is computational simplicity (easy implementation) and adaptive interface with standard video.

Several authors have proposed "foveated" image and video compression algorithms which selectively code image data at variable resolutions according to a presumed fixation point [11][23][36]. Since there is no agreed-upon effective criterion embedded in foveated video processing, most research has been done on the concatenation of foveation filtering and video processing, or weighted video processing with respect to foveation points. However, using the FSNR, it is possible to measure performance improvement objectively and process video algorithms without any coordinate (inverse coordinate) transform. Also, it is possible to avoid additional aliasing noise. The FSNR implies quality assessment over curvilinear coordinates without using any coordinate transformation, which means that maximizing the FSNR over cartesian coordinates leads to optimization over curvilinear coordinates. Another main advantage is that it is possible to reduce the computational overhead by utilizing the non-uniform property of the foveated image without reducing the image size, as shown in [23]. Based on the above motivations, we demonstrate high performance foveated processing algorithms in the following sections.

III. Foveated compression protocols

Fig. 2 shows an example of end-to-end foveated visual communication systems and the foveation point flow. The gray marked modules indicate the modules optimized using the information. In the block diagram, the regular video sequence becomes a foveated video sequence after passing through the foveation filtering. Foveation points are decided using an eye tracker.

Knowledge of the foveation points in advance makes it possible to significantly increase compression performance. As in, for example [37], a QP can be chosen for an assumed gaze point; then, the quantization step size is increased as the inverse of visual sensitivity. A non-uniform quantization algorithm then assigns more bits to the foveated region(s). If the decoder has stored the QP pattern with respect to the selected foveation points, it is not necessary to compress the differential values of the quantization parameters (DQUANT). In other words, if the encoder sends a minimum QP with a macroblock, including a foveation point, then the decoder can reconstruct the quantization pattern over the frame according to the protocol. In addition, the protocol can be also used for motion compensation. When the foveation points vary along the temporal axis, degradation in the performance of the motion compensation algorithm can be ameliorated by increasing the temporal correlation, using knowledge of the movements of the foveation points: both encoder and decoder create foveated sequences (e.g., via foveation filtering) using the updated foveation points over the coding loop. Using this protocol, it is possible to increase the coding performance considerably.

In this paper, we assume that the encoding system receives a series of foveation points using a state-of-the-art eye tracker or equivalent effective apparatus. The coordinates of the foveation point(s) is then sent from the encoder to the decoder; the decoder uses the information to reconstruct QPs, to effect foveation filtering in the motion compensation, and to perform video postprocessing. Therefore, the main goal is to implement efficient video coding algorithms by reducing the computational overhead and by increasing the apparent visual quality relative to the data volume.

IV. Globally versus locally visualized video/image coding

Video/image coding schemes can be divided into two categories: those that are assumed to be globally visualized, with no preferred points of interest, and those that are assumed to be locally visualized, where it is assumed that there are space/time points or regions of higher interest,

e.g., more likely to be fixated upon, more information-rich, etc. The majority of algorithms fall into the first category. In globally visualized coding, the main issue is to improve the global picture quality. Although it is intended that human viewers will be the ultimate consumers of the visual product, the putative gaze points are tacitly assumed uniformly distributed over each frame. Globally-applied quality metrics have been most commonly used to assess the perceptual success of these algorithms, e.g., the MSE and PSNR. The simplicity of the MSE allows it to be embedded into video processing modules such as motion estimation or rate control algorithms. However, the MSE and PSNR correlate only poorly with subjective quality measurements.

In locally visualized video coding algorithms, the human eye is assumed to focus on particular objects or local image regions of high interest. This category of compression algorithms can be further divided into two subcategories: Region-Of-Interest (ROI) coding and foveated video coding. ROI coding has been substantially developed for video teleconferencing applications. It attempts to improve visual quality by specifically allocating more bits to the region of highest interest, viz., the human face. For example, facial features can be detected and tracked using elliptical face and rectangular eyes-nose-mouth models [38], motion, head shape, average skin color [39] and global motion estimation and edge extraction [40]. In algorithms of these types, specific types of interesting objects are automatically detected using pre-defined models.

Foveated video coding has been introduced as an efficient technique for low bit rate visual communications [1][23][36]. Optimal rate control can also be obtained, given an adequate visual quality metric, such as the FSNR developed in [1]. The FSNR has been successfully used to allocate coding resources optimally to control rate in foveated video sequences. Since the FSNR is easily embedded in video processing modules, it is possible to objectively maximize visual quality by applying non-uniform quantization based on optimal rate control criteria.

Table I shows a coding scheme performance comparison where the following abbreviations are used: PRR (perceptual redundancy reduction using preprocessing), CC (computational complexity), PPE (preprocessing error), CE (coding error), pre-pr (pre-processing), rate-c (rate control), OEC (optimal error control) for reducing PPE and CE based on visual system modelling, OP (optimal), NOP (non-optimal).

V. Efficient Foveation Filtering

Foveation filtering is a method of producing a variable-resolution image, with resolution falling smoothly away from a specified foveation point or points. This is accomplished by passing the

TABLE I

PERFORMANCE COMPARISON

	globally visualized coding		locally visualized coding	
coding scheme	regular	global perceptual	ROI	foveated
compression gain	none	pre-pr	pre-pr/rate-c	pre-pr/rate-c
PRR	none	small	medium	none $\sim$ large
CC	small	large	small	small
target rate	middle/high	middle/high	low/middle	very low/low
operation	real-time	non real-time	real-time	real-time
OEC(PPE/CE)	NOP/NOP	OP/NOP	NOP/NOP	OP/OP

image through a bank of low-pass filters of variable cutoff frequency, and by spatially selecting the filter cutoff to be used at each coordinate according to the resolution fall-off criterion. Ideally, a foveation filtering algorithm would be supplied with the ability to specify an arbitrary, continuously-varying lowpass cutoff frequency, but this presents insurmountable computational difficulties. Instead, a finite, discrete set of filters must be implemented. The values of the filter cutoffs can be obtained by using an exponential model of the spatial sampling grid of the fovea in the human retina, in conjunction with an assumed viewing distance, and by choosing the model parameters such that the human eye would be unable to discriminate the foveated image from the original, assuming a given fixation point in the image being observed.

A. Local bandwidth acquisition

The visual angle between the fovea and a point on the retina with respect to the nodal point of the optics in the human eye is defined as *eccentricity*. Fig. 3 (a) shows the viewing parameters: the image size i_p (pixels), the image size i_d (units of length), the distance v_d from the viewer to the image (units of length), and the distance $d_{\mathbf{x}}$ (pixels) from the foveation point to a pixel at $\mathbf{x} = (x_1, x_2)$ (pixels). The eccentricity $e_{\mathbf{x}}$(degrees) is given at the point $\mathbf{x}$ by $e_{\mathbf{x}} = \tan^{-1}(i_d d_{\mathbf{x}}/v_d i_p)$.

We measure the maximum detectable frequency $f_{d_{\mathbf{x}}}$(cycles/degree) at $e_{\mathbf{x}}$ by using the model $f_{d_{\mathbf{x}}} = \gamma/(e_{\mathbf{x}} + \eta)$ where γ and η are parameters that control the spatial frequency decay [41]. Fig. 3 (b) shows the detectable frequency $f_{d_{\mathbf{x}}}$ based on the same shape as the photoreceptor density profile and the human visual model where $\gamma = 18$, $\eta = 0.2$, $v_d = 30$ cm and $i_d = 9$ cm.

Using a conversion factor $\beta_{\mathbf{x}}$, $f_{d_{\mathbf{x}}}$ (cycles/degree) is converted into $f_{p_{\mathbf{n}}}$ (cycles/pixels). Then,

$f_{p_\mathbf{n}}$ is thresholded at the maximum normalized frequency 0.5: $f_{p_n} = \min[\beta_\mathbf{x} f_{d_\mathbf{x}}, 0.5]$ and at a minimum bandwidth f_{min} (here, we set it to 0.07): $f_{p_n} = \max[f_{p_n}, f_{min}]$. Suppose that the foveation point is in the middle of a CIF image. Then, the local bandwidth f_{p_n} is plotted over horizontal pixels in Fig. 3 (c) and over eccentricity in Fig. 3 (d) according to viewing distance.

B. Foveation design

Once the local spatial frequency f_{p_n} has been obtained, a foveated image is obtained by applying a *foveation filter* which consists of a bank of lowpass filters. The cutoff frequencies of the low pass filters are denoted f_{p_n}. To indicate the n^{th} pixel, we use the subscript n or the position vector $\mathbf{n} = (n_1, n_2)$. From a discrete original image $I(\mathbf{m})$, the foveated image $\tilde{I}(\mathbf{n})$ can be obtained by $\tilde{I}(\mathbf{n}) = L^*[I(\mathbf{m}), f_{p_n}]$ where $\mathbf{m} = (m_1, m_2)$ is the position vector and $L^*(., f_{p_n})$ is an ideal lowpass filter with the cutoff frequency f_{p_n}.

According to the Parseval's theorem, the total energy in a discrete signal is equal to the integral of its energy-density spectrum. Let $h^*(i)$ be an ideal lowpass filter with cutoff frequency ω_c, and let $H^*(e^{j\omega})$ be the Fourier transform of $h^*(i)$. Then,

$$\sum_{i=-\infty}^{\infty} h^{*2}(i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H^*(e^{j\omega})|^2 d\omega = \omega_c/\pi$$

where $H^*(e^{j\omega}) = 1$ for $|\omega| < \omega_c$ and $H^*(e^{j\omega}) = 0$ for $\omega_c < |\omega| \leq \pi$. The impulse response of $h^*(i)$ becomes $h^*(i) = \frac{\omega_c}{\pi}\, sinc(\omega_c i)$, $-\infty \leq i \leq \infty$. If ω_c is changed according to each pixel such as $w_c = 2\pi f_{p_n}$, then $H^*(., f_{p_n})$ becomes the ideal low pass filter $L^*(., f_{p_n})$ for creating foveated images.

In a practical implementation, the filter length is finite. Let $h(i)$ and $H(e^{j\omega})$ be the Fourier transform pair of a low pass filter with filter length N. Then, the total energy of the error signal between $h^*(i)$ and $h(i)$ becomes

$$e = \sum_{i=-\infty}^{\infty} [h^*(i) - h(i)]^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H^*(e^{j\omega}) - H(e^{j\omega})|^2 d\omega. \quad (1)$$

In order to minimize e, $h(i)$ must be $h(i) = h^*(i)$, if $-N/2 \leq i \leq N/2$, and $h(i) = 0$, otherwise. The error ratio τ relative to the total energy becomes $\tau = e\pi/\omega_c$. Given e, τ is inversely proportional to the cutoff frequency ω_c. As ω_c is decreased, the value N must be increased to achieve a small error ratio. A minimum N should be used such that τ falls below a given constant. Then, the filter length varies according to position with reference to the fixed value

τ, because the local bandwidth is a function of position. Thus, the filter length N becomes a position-dependent variable N_n. In Fig. 4, the filter length N is adaptively changed according to the cutoff frequency, when the upper bounds of τ are 85 %, 90 % and 95 %.

C. Separable even symmetric filters

To smooth the frequency response of the low pass filter, a Hamming window $w(i_1, i_2)$ is employed. Then, the designed foveation filter $l(i_1, i_2)$ can be expressed as $l(i_1, i_2) = h(i_1, i_2)\, w(i_1, i_2)$. Since the cutoff frequency depends on position, $l(.)$ is also a function of n, e.g., $l_n(.)$. Then, the foveated image can be obtained as

$$\tilde{I}(n_1,\, n_2) = \sum_{i_1=-N_n/2}^{N_n/2} \sum_{i_2=-N_n/2}^{N_n/2} I(n_1 - i_1,\, n_2 - i_2) l_n(i_1,\, i_2) \tag{2}$$

To reduce the number of multiplications, we use a separable even-symmetric low-pass filter $l(i_1, i_2)$:

$$l_n(i_1, i_2) = \begin{cases} l_n(i_1) l_n(i_2), & \text{if} - N_n/2 \le i_1,\ i_2 \le N_n/2 \\ 0, & \text{otherwise} \end{cases}$$

and $l_n(i_1) = l_n(-i_1)$. Then,

$$\tilde{I}(n_1,\, n_2) = \sum_{i_1} l_n(i_1) \sum_{i_2} I(n_1 - i_1,\, n_2 - i_2) l_n(i_2).$$

For the separable even-symmetric filter, the number of operations is reduced to $2 * (N_n/2 + 1)$ for additions and multiplications for each pixel.

D. Circularly symmetric filters

By using circularly symmetric filters, it is possible to obtain a more symmetric frequency response for each foveation filter/bandwidth. To reduce the number of multiplications, we exploit the octal symmetry of circularly symmetric filters $l_n(i_1,\, i_2) = l_n(\pm i_1,\, \pm i_2) = l_n(\pm i_2,\, \pm i_1)$. Then, (2) becomes

$$\tilde{I}(n_1,\, n_2) = \sum_{\substack{i_1,\, i_2 \neq 0 \\ i_1 = 0,\, 1,\, \ldots,\, N_n/2 \\ i_2 = 0,\, 1,\, \ldots,\, i_1}} l_n(i_1,\, i_2) \Big[I(n_1 \pm i_1,\, n_2 \pm i_2) + I(n_1 \pm i_2,\, n_2 \pm i_1) \Big]$$

$$+ I(n_1,\, n_2) l_n(0,\, 0) \tag{3}$$

Using (3), it is possible to compute the number of additions and multiplications that are required to implement this circularly symmetric filter. The number of operations is 7 $(N_n/2 + 1)(N_n/2 +$

$2)/2-6$ for additions and $(N_n/2+1)(N_n/2+2)/2-1$ for multiplications. Thus, due to the octal symmetry, the number of multiplications required for implementing circular symmetric filters has been reduced by an approximate factor of 8 for $N_n \rightarrow \infty$.

E. Simulation results

Suppose that the distribution of foveation points is probabilistic with probabilities described by a Gaussian probability density function (PDF) falling away from the center point in the image. Let H/V be the number of pixels in each horizontal/vertical line. The PDF is

$$p(n_1, n_2) = \alpha e^{-2\pi^2\sigma^2 r^2/i_p^2} \tag{4}$$

where $i_p = max[H,\, V]$, $r = \sqrt{(n_1 - i_p/2)^2 + (n_2 - i_p/2)^2}$, $1 \leq n_1 \leq H$, $1 \leq n_2 \leq V$, and α is a constant. Usually the half-peak radius r_c is a factor to select σ by $e^{-2\pi^2\sigma^2 r_c^2/i_p^2} = 1/2$ giving $\sigma = (\frac{i_p}{\pi r_c})\ \sqrt{log\sqrt{2}} \approx 0.19\ (\frac{i_p}{r_c})$. The value α is decided in order that the sum of $p(n_1, n_2)$ for all pixels is equal to 1. Then, the average number of operations is

$$\bar{O}_T = \sum_{n_1=1}^{H} \sum_{n_2=1}^{V} p(n_1, n_2) O(N_n) \tag{5}$$

where $O(N_n)$ is the total number of operations when the foveation point is at the n^{th} pixel.

Table II shows the average number of multiplications per pixel. As the error ratio τ is reduced, the number of operations rapidly increases. When $\tau = 0.05$, the number of operations incurred by circularly symmetric filtering is much larger than with separable even symmetric filtering. However, when $\tau = 0.15$, the number of operations is similar for both methods. When σ is smaller, the foveation points are more widely distributed relative to the center point of the image. When the foveation point is located away from the center point, the region associated with small local bandwidths increases in size so that, overall, the number of operations is increased. Thus, the number of operations for $\sigma = 0.38$ is larger than for $\sigma = 0.57$.

Fig. 5 (a) shows the original *lena* image, and Figs. 5 (b) to (d) show the foveated images according to the filter length and filter coefficients. In Fig. 5 (b), the filter length N is fixed to 31 and circularly symmetric filters are used. In Fig. 5 (c) and (d), the filter length is adaptively changed according to the local bandwidth for $\tau = 0.1$, and both types of filters are used. It can be seen that the visual quality of the foveated images are similar to each other, although the computational complexity is substantially reduced by regulating the number of filter lengths and by using separable even symmetric filters.

VI. Efficient motion estimation

Block matching algorithms have been widely used to estimate the displacement of moving objects in video streams [42][43]. In the MPEG and H.263 implementations, motion vectors are chosen over a search window using a MSE (mean square error) or MAD (mean absolute distortion) criterion. Since the complexity of the MAD is lower than that of the MSE, the MAD has been more widely used. Here, one of hierarchical block matching algorithms (HBMAs) [43][44][45] is employed and applied to the foveated video coding based on a foveal weighting factor.

Given original and compressed versions of a foveated image, both can be mapped into curvilinear coordinates as in Fig. 1 (d); then, the MSE can be measured over these coordinates. Let the local bandwidth be proportional to the local sampling density, which corresponds to the Nyquist sampling rate in the curvilinear coordinates. Then, the *Jacobian* of the coordinate transformation can be approximated as the square of the local bandwidth, which becomes a pixel-to-pixel mapping ratio. Thus, visual quality can be assessed over virtual curvilinear coordinates. The FMSE (foveal mean square error) is defined as:

$$\mathrm{FMSE} = \frac{1}{\sum_{n=1}^{N} f_n^2} \sum_{n=1}^{N} \left[a(\mathbf{x}_n) - b(\mathbf{x}_n)\right]^2 f_n^2, \tag{6}$$

and the foveal PSNR (FPSNR) is

$$\mathrm{FPSNR} = 10\log_{10} \frac{\max[a(\mathbf{x}_n)]^2}{\mathrm{FMSE}} \tag{7}$$

where f_n is the local bandwidth at the n^{th} point and $b(\mathbf{x}_n)$ is a compressed version of an original frame $a(\mathbf{x}_n)$ or a foveated frame $a(\mathbf{x}_n)$ [1]. Similarly, the foveal MAD (FMAD) is defined by

$$\mathrm{FMAD} = \frac{1}{\sum_{n=1}^{N} f_n^2} \sum_{n=1}^{N} \left|a(\mathbf{x}_n) - b(\mathbf{x}_n)\right| f_n^2, \tag{8}$$

Fig. 1 (d) shows the uniform version of the foveated image Fig. 1 (b). The reducing ratio increases towards the periphery. For a given motion compensation error at a pixel, the human eye tends to be sensitive in proportion to the mapping ratio f_n^2 of the pixel over the spatial domain. Motion vectors are found using a full search over the uniform version of the image Fig. 1 (d) which corresponds to an adaptive sub-sampled search over the foveated image Fig.

1 (b). The main motivation of this algorithm is to minimize motion compensated errors near the foveation (and presumed fixation) point using quality motion estimation, and to reduce the computational complexity using sparse motion estimation away from the foveation point. To achieve this goal, we present an HBMA, wherein the subsampling rate is adaptively modified according to the average of the local bandwidths in a macroblock.

A. HBMA using subsampling

We will use the following notation.

H	the number of pixels along the horizontal axis
V	the number of pixels along the vertical axis
$I_k(i,j)$	the luminance at coordinate (i,j) of frame k ($1 \le i \le H$, $1 \le j \le V$)
(m,n)	the coordinate of the macroblock whose top-left corner is located at (m,n)
(v_x, v_y)	the horizontal and vertical displacement
W	the window size for searching motion vectors
$f_{i,j}$	the local bandwidth at the point (i, j) of frame k

The FMAD between the current macroblock and the motion compensated macroblock is

$$\text{FMAD} = \left[\sum_{i,j=0}^{M-1} f_{m+i,n+j}^2\right]^{-1} \sum_{i,j=0}^{M-1} |I_k(m+i,n+j) - I_{k-1}(m+i+v_x, n+j+v_y)| f_{m+i,n+j}^2 \quad (9)$$

where $M \times M$ is the number of pixels in a macroblock ($M = 16$ in the H.263/MPEG standard). Then, the motion vector (v_x^*, v_y^*) which minimizes the FMAD is found by $(v_x^*, v_y^*) = \text{argmin}_{(v_x,v_y)}$ FMAD (v_x, v_y) where $-W/2 \le v_x, v_y \le W/2 - 1$.

Let $\bar{f}_{m,n}$ be the average local bandwidth over a macroblock whose upper left corner position is (m, n). The magnitude of the local bandwidth generally varies smoothly, so assume that $\bar{f}_{m,n} \simeq f_{m+i,n+j}$ $(0 \le i, j \le M-1)$. Since the maximum normalized frequency in the discrete domain is 0.5, the subsampling factor $d_{m,n}$ is approximately $d_{m,n} = \lfloor 0.5/\bar{f}_{m,n} \rfloor$. The FMAD in (9) is represented by

$$\text{FMAD} \simeq \frac{1}{\lfloor M/d_{m,n} \rfloor^2} \sum_{i,j=0}^{\lfloor M/d_{m,n} \rfloor - 1} |I_k(m+d_{m,n}i, n+d_{m,n}j) - I_{k-1}(m+d_{m,n}i+\tilde{v}_x, n+d_{m,n}j+\tilde{v}_y)| \quad (10)$$

where $(\tilde{v}_x, \tilde{v}_y) = (\alpha d_{m,n}, \beta d_{m,n})$, $\alpha \in \mathbf{N}$, $\beta \in \mathbf{N}$. Then, the motion vector is obtained by (v_x^*, v_y^*) $= \text{argmin}_{(\tilde{v}_x,\tilde{v}_y)}$ FMAD$(\tilde{v}_x, \tilde{v}_y)$ where $-W/2 \le \tilde{v}_x, \tilde{v}_y \le W/2 - 1$. Since $\lfloor M/d_{m,n} \rfloor^2$ is a constant,

minimizing the FMAD is equivalent to minimizing the MAD.

Fig. 6 shows the block diagram for the HBMA. First, the subsampling rate is set to be $d_{m,n^1} = \lfloor 0.5/\bar{f}_{m,n} \rfloor$. Then find $(v^*_{x_1}, v^*_{y_1})$ over the subsampled version of the search area where 1 indicates the first level. In level 2, the subsampling rate is reduced to $d_{m,n^2} = d_{m,n^1} - 1$. After obtaining the subsampled block from the current macroblock, the block is matched with 9 neighborhood blocks with reference to $(v^*_{x_1}, v^*_{y_1})$. Among nine motion vectors $(v^*_{x_1}+[\pm 1,0], v^*_{y_1}+[\pm 1,0])$, the one that minimizes the MAD is chosen. Similar to level 2, the motion vector $(v^*_{x_3}, v^*_{y_3})$ in level 3 is found. This motion estimation process is continued until the subsampling rate becomes 1. In general, the MAD of the k^{th} level is

$$\mathrm{MAD}_k = \sum_{i,j=0}^{\lfloor M/d_{m,n^k} \rfloor - 1} |I_k(m + d_{m,n^k} i, n + d_{m,n^k} j) - I_k(m + d_{m,n^k} i + v_{x_k}, n + d_{m,n^k} j + v_{y_k})| \quad (11)$$

where d_{m,n^k}=$d_{m,n} - k + 1$, $v^*_{x_{k-1}} - 1 \le v_{x_k} \le v^*_{x_{k-1}} + 1$ and $v^*_{y_{k-1}} - 1 \le v_{y_k} \le v^*_{y_{k-1}} + 1$.

Suppose that the total number of levels for the current macroblock is L. Then, the number of operations needed to find a motion vector is

$$o(L) = \lfloor \frac{M}{2^{L-1}} \rfloor^2 \lfloor \frac{W}{2^{L-1}} \rfloor^2 + 9 \sum_{l=1}^{L-1} \lfloor \frac{M}{2^{L-l-1}} \rfloor^2. \quad (12)$$

The range of the subsampling rate for the above HBMA becomes $1 \le d_{m,n^k} \le d_{m,n}$. At the cost of possibly increasing motion compensation errors, the number of operations can be reduced by increasing the subsampling rate such as $2 \le d_{m,n^k} \le d_{m,n} + 1$. Then, the number of operations is reduced:

$$o(L) = \lfloor \frac{M}{2^{L}} \rfloor^2 \lfloor \frac{W}{2^{L}} \rfloor^2 + 9 \sum_{l=1}^{L} \lfloor \frac{M}{2^{L-l}} \rfloor^2. \quad (13)$$

Since the total number of macroblocks is $N_m = VH/M^2$, the total operations for each frame become $O_T = \sum_{p=0}^{N_m} o(L_p)$ where L_p is the number of levels for the p^{th} macroblock. For full-search motion estimation, the number of operations becomes $O_T = N_m M^2 W^2$.

B. Simulation results

For the simulation, 30 CIF "mobile" frames are used. Fig. 1 (b) shows the first frame of the sequence, where the foveation point is indicated by the mark "x" on the ball. To compare the performances of these, we employ the following three methods: • *FS*: full search motion estimation, • *HBMA 1*: HBMA using the operation (12), • *HBMA 2*: HBMA using the operation (13).

Table III shows the percentage of the level and the operation when the foveation point distribution $\sigma = 0.57$ in (4). In FS, the operation number for each macroblock is 246016. In Table III, the average operation for $HBMA\ 1$ ($HBMA\ 2$) is 37682 (5242). Thus, the ratio of operations between the methods becomes $FS : HBMA\ 1 : HBMA\ 2 = 46.93 : 7.19 : 1$. When $\sigma = 0.38$, this ratio becomes $FS : HBMA\ 1 : HBMA\ 2 = 48.18 : 7.01 : 1$.

Fig. 7 (a) and (b) show the MAD and the FMAD according to the motion estimation methods for regular and foveated video sequences. By exploiting the non-uniform resolution, the performance of the $HBMA$ method approaches the optimal performance obtained by FS. So, motion compensated errors are very effectively reduced in the foveated video sequence. In the regular video, the $HBMA$ methods produce more motion compensation errors than the FS method. In Fig. 7, the MAD/FMAD difference by the above methods is relatively small in the foveated video sequence compared to the regular video sequence. Since high frequencies are eradicated away from the foveation point, motion vectors of adjacent macroblocks are highly correlated. So, it is possible to improve on the number of bits needed to compress the difference of motion vectors.

Table IV shows the average value of MAD and FMAD for regular and foveated video sequences. To create the foveated sequences, separable even symmetric filters and circularly symmetric filters are used. In Table IV, motion compensated errors using circularly symmetric filters are effectively reduced since high frequency components are removed. In particular, the FMAD under $HBMA\ 1$ is the almost same as the FMAD under FS, i.e., the visual quality of the motion compensated frames is nearly the same as when the human eye gazes at the foveation point.

VII. Motion compensation

Motion compensation techniques have been used to reduce temporal redundancy in video compression algorithms. Motion estimation is a required precursor to motion compensation. In practice, motion estimation refers to a search for motion displacements using a block matching algorithm. Motion compensation refers to taking one of those motion vectors to improve the efficacy of a video processing algorithm. If the correlation between the current frame and the reference frame is increased (P- or I-frame in MPEG), then motion compensation errors will be reduced.

When a foveation point is added or dropped, the local bandwidth also varies so that the temporal correlation is decreased between two frames. Depending on the change in the temporal

correlation, motion compensation errors will also fluctuate. We introduce a motion compensation algorithm for foveated video, which reduces motion compensation errors by increasing the temporal correlation. The filter operation is performed at the encoder as well as at the decoder, within the coding loop via the foveation protocol.

A. Local bandwidth for multiple foveation points

In [23][46], the coordinate of each point over the transform domain was calculated with respect to the location of multiple foveation points. Then, the video processing was done over the new coordinate domain. In our approach, the local bandwidth for multiple foveation points is obtained from the location information. Using the local bandwidth, a foveated image with multiple foveation points is obtained by using foveation filtering over the original coordinate domain.

The local bandwidth for a single foveation point can be used to obtain the local bandwidths when there are multiple foveation points. Assume that there exist P different foveation points in a picture. Let $f^k_{p_{\mathbf{x}}}$ be the local bandwidth at pixel $\mathbf{x}$ with reference to the k^{th} foveation point. Therefore, the optimal local bandwidth $f_{p_{\mathbf{x}}}$ is the maximum value of $f^k_{p_{\mathbf{x}}}$ ($1 \leq k \leq P$) :

$$f_{p_{\mathbf{x}}} = \max(f^1_{p_{\mathbf{x}}}, \ f^2_{p_{\mathbf{x}}}, .. \ f^P_{p_{\mathbf{x}}}) \tag{14}$$

Fig. 8 (a) shows a foveated image with multiple foveation points, whose local bandwidth in Fig. 8 (b) is decided by (14).

B. Motion compensation for foveated video

To describe the algorithm, the following notation are used : I^c – the current frame; p^c_i – the i^{th} foveation point of I^c; $\mathbf{F}^c$ – the set of current foveation points, $p^c_i \in \mathbf{F}^c$; I^p_d – the coded reference frame; I^c_m – the motion compensated frame; p^p_i – the i^{th} foveation point of I^p_d; $\mathbf{F}^p$ – the set of foveation points in I^p_d, $p^p_i \in \mathbf{F}^p$; $I(\mathbf{F})$ – the foveated image with the set $\mathbf{F}$, $p_i \in \mathbf{F}$.

The set which consists of such commonly shared foveation points in I^c and I^p_d is denoted $\mathbf{F}^s = \mathbf{F}^c \cap \mathbf{F}^p$. When foveation points move to other objects or are newly added to the set $\mathbf{F}^c$, the set of such foveation points is denoted $\mathbf{F}^a = \mathbf{F}^c - \mathbf{F}^s$. The set which consists of removed foveation points, i.e., $p^p_i \in \mathbf{F}^p$ and $p^p_i \notin \mathbf{F}^c$ is denoted as $\mathbf{F}^d = \mathbf{F}^p - \mathbf{F}^s$. If $\mathbf{F}^a \neq \phi$ or $\mathbf{F}^d \neq \phi$, the local bandwidth of $I^c(\mathbf{F}^c)$ and $I^p_d(\mathbf{F}^p)$ is different around the foveation points, $p^c_i \in \mathbf{F}^a$ or $p^p_i \in \mathbf{F}^d$.

In general, $I^p_d(\mathbf{F}^p)$ is used as a reference frame in computing motion estimates for the current

foveated frame $I^c(\mathbf{F}^c)$. For a given motion estimation algorithm ME, the motion vectors $\mathbf{v}$ are obtained by

$$\mathbf{v} = ME(I^c(\mathbf{F}^c), I_d^p(\mathbf{F}^p)) \tag{15}$$

Using motion compensation MC with $\mathbf{v}$, the motion compensated frame $I_m^c(\mathbf{F}^c)$ is given by $I_m^c(\mathbf{F}^c) = MC(I_d^p(\mathbf{F}^p), \mathbf{v})$. Then, the prediction error e_m becomes $e_m = I^c(\mathbf{F}^c) - I_m^c(\mathbf{F}^c)$. The prediction error e_m is large around $p_i^c \in \mathbf{F}^a$ or $p_i^p \in \mathbf{F}^d$ due to the difference of the local bandwidth between $I^c(\mathbf{F}^c)$ and $I_d^p(\mathbf{F}^p)$.

In order to maintain a high temporal correlation and reduce prediction errors, the local bandwidth must be similar between $I^c(\mathbf{F}^c)$ and $I_d^p(\mathbf{F}^p)$. If the motion of foveated objects is slow between $I^c(\mathbf{F}^c)$ and $I_d^p(\mathbf{F}^p)$, and the observers eye successfully follows the foveation point along consecutive frames, then the local bandwidth of $I_d^p(\mathbf{F}^p - \mathbf{F}^d)$ becomes similar to that of $I^c(\mathbf{F}^c - \mathbf{F}^a)$ with respect to the foveation points. Thus, $I_d^p(\mathbf{F}^p - \mathbf{F}^d)$ is used as the reference frame for $I^c(\mathbf{F}^p)$ instead of $I_d^p(\mathbf{F}^p)$. Also, $I(\mathbf{F}^c - \mathbf{F}^a)$ can be used for motion estimation rather than $I(\mathbf{F}^c)$. Then, motion vectors can be obtained by $\mathbf{v} = ME(I^c(\mathbf{F}^c - \mathbf{F}^a),\ I_d^p(\mathbf{F}^p - \mathbf{F}^d))$. The motion compensated frame $I_m^c(\mathbf{F}^c - \mathbf{F}^a)$ becomes $I_m^c(\mathbf{F}^c - \mathbf{F}^a) = MC(I_d^p(\mathbf{F}^p - \mathbf{F}^d), \mathbf{v})$.

C. Simulation results

In the foveated image Fig. 8 (a), there are three foveation points with the set { p_1 = (93, 143), p_2 = (168, 96), p_3 = (267, 141) } from left-to-right. In order to test the motion compensation algorithm, we use 30 CIF "News" frames, where the number of foveation points is assumed to be 3 (1 to 10 frames where $\mathbf{F}^c = \{p_1, p_2, p_3\}$), 1 (11 to 20 frames where $\mathbf{F}^c = \{p_1\}$) and 2 (21 to 30 frames where $\mathbf{F}^c = \{p_2, p_3\}$). When the eleventh frame is coded, $\mathbf{F}^c = \{p_1\}$ and $\mathbf{F}^p = \{p_1, p_2, p_3\}$ which yield $\mathbf{F}^s = \{p_1\}$, $\mathbf{F}^a = \phi$, $\mathbf{F}^d = \{p_2, p_3\}$. Similarly, in the twenty-first frame, $\mathbf{F}^c = \{p_2, p_3\}$, $\mathbf{F}^p = \{p_1\}$, $\mathbf{F}^s = \phi$, $\mathbf{F}^a = \{p_2, p_3\}$ and $\mathbf{F}^d = \{p_1\}$. In order to improve the temporal correlation, $I_d^p(\mathbf{F}^s)$ is used instead of $I_d^p(\mathbf{F}^p)$. In the simulation, we use the current frame $I^c(\mathbf{F}^c)$ instead of $I^c(\mathbf{F}^c - \mathbf{F}^a)$. Thus, $\mathbf{v}$ is obtained by $\mathbf{v} = ME(I^c(\mathbf{F}^c), I_d^p(\mathbf{F}^s))$.

In Table V, the prediction errors are measured by the MAD and the FMAD when the foveation points are changed. Using the proposed algorithm, we can reduce the prediction errors, which also leads to improvement in the performance of motion compensation.

Fig. 9 shows another example, when foveation points are changed from $\{p_1, p_2, p_3\}$ to $\{p_1\}$ in the 11^{th} frame and to $\{p_3\}$ in the 21^{th} frame. In the figure, *prop. MC*(*reg. MC*) indicates

the proposed (regular) method. The improvement in quality in the later frames is due to the improved quality in the 11^{th} and 21^{th} frames.

VIII. Rate control

A. Review of optimal rate control

Rate control methods are typically optimized by minimizing the overall distortion $D(\vec{Q})$ of a frame subject to the rate constraint $R(\vec{Q}) \leq R_T$ where $\vec{Q}$ is the vector of QPs in the frame, $R(\vec{Q})$ is the number of generated bits using $\vec{Q}$ and R_T is the number of target bits assigned to the frame. The vector $\vec{Q}$ is expressed by $\vec{Q} = (q_1, q_2, ..., q_M)$ where q_k is the QP of the k^{th} macroblock and M is the number of macroblocks in the frame. Usually, the MSE has been used as a quality metric to measure the distortion $D(\vec{Q})$. However, it is not suitable for measuring distortion in spatially localized (foveated) video. For foveated video, the FMSE can be embedded in the rate control algorithm and utilized as a quality metric, like the MSE is used in standard algorithms, to find an optimal vector of QPs. In [2], the visual quality of reconstructed foveated images was demonstrated to improve by minimizing the FMSE embedded in a rate control algorithm.

In H.263 video coding, the variation of QPs results in changing the quantization mode and the coding type of the following macroblocks. The macroblock type is combined with the coded block pattern for chrominance. When the QP is changed, the code length is increased to $2 - 8$ bits. Due to the code length dependency along consecutive macroblocks, the rate-distortion curve cannot be independently obtained at each macroblock. In [2], a constant QP was decided for macroblocks whose average local bandwidth was less than a threshold value. For the remaining macroblocks, QPs were chosen based on the Lagrangian cost. However, even though the rate-distortion model is used, it is necessary to code several times to obtain an optimal Lagrange multiplier λ^*.

B. Circularly symmetric quantization parameter set

As an approach to optimal performance, the QP of each macroblock is set to be inversely proportional to the average value of the local bandwidth. Since the local bandwidth is circularly symmetric with respect to a foveation point (in the sense of foveation filtering), and is minimum at the foveation point, then the spatial distribution of QPs is also circularly symmetric. The QP at the macroblock containing the foveation point is set to a minimum value Q_m, and the rest of the QPs are decided using the average value of the local bandwidth and Q_m.

The QP of the k^{th} macroblock is given by

$$q_k = \frac{Q_m}{[2\bar{f}_k]^n} \tag{16}$$

where $\bar{f}_k$ is the average local bandwidth of the k^{th} macroblock and $n \in \mathbf{N}$. If $\bar{f}_k$ is the maximum discrete frequency 0.5, then $q_k = Q_m$. This method can be generalized to an image with multiple specified foveation points. For multiple foveation points, the local bandwidth at each pixel is obtained by (14). Then, q_k is also obtained using (16). In real systems, the encoder sends Q_m and the coordinates of the foveation points to the decoder. The decoder can reconstruct the quantization parameter q_k using (16). Using this protocol, it is possible to save the number of bits required to represent DQUANT.

C. Simulation results

In the simulations, the CIF foveated "Mobile" and "News" sequences are compressed using the H.263 algorithm, where each sequence consists of 60 frames. The reference frame rate is 30 (frames/sec.), and the target frame rate is 10 (frames/sec.) for the "News" sequence and 15 (frames/sec.) for the "Mobile" sequence. In order to measure the quality of P pictures, the first reference picture (I picture) is compressed using QP = 12. The performance comparison is conducted via four rate control methods.

- *const_q :* A constant QP is used, where $12 \leq$ QP ≤ 31.
- *linear :* A minimum QP Q_m is used, where $7 \leq Q_m \leq 31$. The QP is decided by (16) with $n = 1$.
- *square :* A minimum QP Q_m is used, where $7 \leq Q_m \leq 31$. The QP is decided by (16) where $n = 2$.
- *optimal :* Optimal rate control for minimizing the FMSE using Lagrange multiplier where the number of generated bits due to the QP difference is not considered.

The FPSNR of the method *optimal* becomes the upper bound of other rate control methods. The PSNR/FPSNR for the "Mobile" and "News" sequences are shown in Fig. 11. The horizontal axis indicates the average number of generated bits per each frame. The PSNR in the method *const_q* is higher than that in other methods and the method *optimal* exhibits the lowest PSNR due to nonuniform quantization. Conversely, the FPSNR in the method *optimal* is the largest compared to other methods. The FPSNR is decreased in the order of *square*, *linear* and *const_q*.

The FPSNR in the method *square* approaches the upper bound compared to other methods. At very low bit rates, the FSNR approaches the optimal performance. For P frames, the performance also improves according to Fig. 11 (d) to (f).

In order to demonstrate the reconstructed visual quality, we compress the "Lena" image using the H.263 video algorithm. In the "Lena" image, the foveation point is assumed to be located between the subjects two eyes. Using the method *const_q* with QP = 25, we obtain the compressed version of the regular image in Fig. 10 (a) where blocking artifacts are apparent. At the equivalent rate, the reconstructed image in Fig. 10 (b) is obtained using the method *square* with $Q_m = 14$. Because of the nonuniform quantization, the visual quality in Fig. 10 (b) is improved relative to Fig. 10 (a). Note that the PSNR is decreased to 0.4 dB and the FPSNR is improved to 1 dB from Fig. 10 (a) to (b). Thus, it is shown that the FPSNR effectively measures localized visual quality over the spatial domain.

Also, we obtain the compressed version of the foveated "Lena" image in Fig. 10 (c) using the method *square* and (d) using the method *linear* at the equivalent rate. The PSNR/FPSNR is increased compared to Fig. 10 (a) and (b) because of the reduction of high frequency components in the hair and edge regions. Using the method *square*, the FPSNR is improved to 0.3 dB compared to the method *linear* for the foveated image. The visual quality in Fig. 10 (c) demonstrates the highest objective quality.

D. Postprocessing

In H.263+ Annex J, the deblocking filter mode is described [47]. In an effort to reduce blocking artifacts, a block edge filter is used within the coding loop and the filtering is performed on 8×8 block edges at both the encoder and the decoder. Since the effect of blocking artifacts tends to be proportional to the magnitude of QP, the bandwidth of lowpass filtering depends on the magnitude of QP.

If the method *square* is used, the magnitude of the QPs is symmetrically increased according to the distance from a foveation point. The bandwidth of lowpass filtering is also symmetrically decreased. Fig. 12 shows the reconstructed images after reducing blocking artifacts from Fig. 10. In the foveated region near the eyes, the original information is well-preserved. Moreover, blocking artifacts are effectively removed from the background regions.

IX. Simulation results for the overall video processing

Fig. 9 shows the performance improvement for the overall video processing. In the simulation, the CIF foveated "News" sequence (30 frames) is compressed, the reference frame rate is 10 (frames/sec.) with two skip pictures, and the target data rate (only for the P frame sequence) is 30 *kbps*. For foveation filtering, circularly symmetric filters are used with the error ratio $\tau = 0.1$. Two motion estimation and three rate control methods are used: FS, $HBMA$ 2 and *const_q*, *square*, *optimal*. The location of foveation points is changed in the 11^{th} and 21^{th} frames.

In Fig. 9, it is shown that the PSNR/FPSNR is improved by increasing the correlation by the loop filtering. In the 11^{th} and 21^{th} frames, the PSNR (FPSNR) is improved by 2 (0.8) dB and 0.8 (0.6) dB respectively. Since the high frequency redundancy is effectively reduced, the MAD/FMAD difference by the motion estimation methods is relatively small compared to the performance difference by the rate control methods.

The PSNR is higher in the order of the *const_q*, *square* and *optimal* methods. Conversely, the FPSNR is improved in the reverse order. Compared to the *const_q* method, the FPSNR in the *square* method is the FPSNR is improved to 0.4 dB. Some of the foveated video results have been demonstrated on the web site [48].

X. Conclusions

In this paper, we developed efficient new video processing algorithms by exploiting the characteristics of nonuniform spatial resolution for foveated image/video compression. In order to achieve the main goal, two major factors were taken into account for algorithm development: foveated image/video analysis over the uniform domain, and foveation protocols. Then, the performance was measured by using objective quality criteria (SNR and FSNR).

After interpreting non-uniform foveated images over the uniform domain, we applied uniform operations to the uniform (coordinate-mapped) foveated image which results in adaptive video processing, i.e., more accurate video processing over high-resolution regions of the foveated image and simplified video processing over the low-resolution background regions. Using this approach to non-uniform video processing, it is possible to reduce the computation redundancy effectively while approaching optimal performance.

For foveation filtering, we reduced the number of operations by using separable even symmetric or circularly symmetric filters. We further reduced the computational complexity by changing

the filter length (number of taps) according to the magnitude of the lowpass cutoff frequencies. Motion estimation was improved by introducing a hierarchical block matching algorithm (HBMA). The subsampling rate is adaptively changed according to the local bandwidth, and the computational burden is reduced up to $47 : 7.2$ for the full search method relative to the proposed method, while finding near-optimal motion vectors. After obtaining the bandwidth for single or multiple foveation points, the temporal correlation is increased, and the motion compensation errors reduced, by postprocessing in a manner consistent with the changing foveation points. To achieve optimal rate control, quantization information is described using a minimum QP at the macroblock containing a foveation point and using the coordinates of the foveation point. The decoder reconstructs the QPs using the foveation protocol. Standard postprocessing is also adapted to remove blocking artifacts from the reconstructed images.

Foveated visual systems have high potential for application in wireless visual communication systems. Currently, a robust channel adaptation over feedback channels has been studied [4] and an error resilience scheme has been introduced by utilizing human foveation [5]. For future foveated multimedia services, the following applications will be investigated: video conferencing, low bit rate movies, video broadcasting, remote surveillance, internet news, and so on. Such multimedia services will be merged with wireless network technology in the near future and will be provided over limited bandwidths. In 3G wireless networks, limited bandwidths are available: $8.8 - 384$ Kbps in UMTS and $9.6 - 153.6$ Kbps in CDMA 3G-1X. In addition, the channel bandwidth is also dynamically changed according to slow fading, multipath fading, mobile velocity, interference and so on. Clearly, by reducing the visual redundancy using foveation, it is possible to absorb those channel dynamics and to deliver much improved visual quality to end-users. Therefore, the interoperation with wireless environments will be a ley feature to be studied in future work.

References

[1] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, pp. 129–132, March 2002.

[2] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Trans. Image Processing*, vol. 10, pp. 977–992, July 2001.

[3] S. Lee and A. C. Bovik, "Very low bit rate foveated video coding for H.263," in *Proc. IEEE ICASSP'99*, (Phoenix, AZ), pp. VI3113 – VI3116, Mar. 1999.

[4] S. Lee, A. C. Bovik, and Y. Y. Kim, "Low delay foveated visual communications over wireless channels," in

Proc. IEEE ICIP'99, (Kobe, Japan), Oct. 1999.

[5] 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 Systems for Signal, Image and Video Technology*, vol. 32, invited paper, special issues on Multimedia Communications, 2002.

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

[7] D. Pollen and S. Ronner, "Visual cortical neurons as localized spatial frequency filters," *IEEE Trans. Syst., Man, Cybern.*, vol. 13, pp. 907–916, Sep. 1983.

[8] R. E. Kronauer and Y. Y. Zeevi, "Reorganization and diversification of signals in vision," *IEEE Trans. Syst., Man, Cybern.*, vol. 15, pp. 91–101, Jan./Feb. 1985.

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

[10] P. G. J. Barten, "Evaluation of subjective image quality with the square-root integral method," *J. Opt. Soc. Amer.*, vol. 7, pp. 2024–2031, Oct. 1990.

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

[12] M. S. Banks, A. B. Sekuler, and S. J. Anderson, "Peripheral spatial vision: limits imposed by optics, photoreceptors, and receptor pooling," *J. Opt. Soc. Amer.*, vol. 8, pp. 1775–1787, Nov. 1991.

[13] E. Peli, "Contrast in complex images," *J. Opt. Soc. Amer.*, vol. 7, pp. 2032–2039, Oct. 1990.

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

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

[16] K. J. Wiebe, *Variable resolution vision: biologically motivated foveal compression and prioritization*. Univ. of Alberta, 1996.

[17] E. L. Schwartz, "Spatial mapping in primate sensory projection : analytic structure and relevance to perception," *Biological Cybernetics*, vol. 25, no. 4, pp. 181–194, 1977.

[18] E. L. Schwartz, "Computational anatomy and functional architecture of striate cortex:spatial mapping approach to perceptual coding," *Vision Research*, vol. 20, no. 4, pp. 645–669, 1980.

[19] G. Sandini and T. Tagliasco, "An anthropomorhpic retina-like structure for scene analysis," *Comput. Graphics and Image Processing*, no. 14, pp. 365–372, 1980.

[20] R. S. Wallace, P. Ong, B. B. Bederson, and E. L. Schwartz, "Space variant image processing," *International Journal of Computer Vision*, vol. 13, pp. 71–90, Sep. 1994.

[21] M. Tistarelli and G. Sandini, "On the advantages of polar and log-polar mapping for direct estimation of time-to-impact from optical flow," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 401–410, April 1993.

[22] W. N. Klarquist and A. C. Bovik, "Fovea: A foveated, multi-fixation, vergent active stereo system for dynamic three-dimensional scene recovery," *IEEE Trans. Robotics and Automation*, 1999.

[23] A. Basu and K. J. Wiebe, "Enhancing videoconferencing using spatially varying sensing," *IEEE Trans. Syst., Man, Cybern. -part A : systems and humans*, vol. 28, pp. 137–148, March 1998.

[24] A. Basu, A. Sullivan, and K. J. Wiebe, "Variable resolution teleconferencing," in *IEEE International Conference on systems, man & cybernetics*, (France), pp. 170–175, Oct. 1993.

[25] J. I. Yellott, "Image sampling properties of photoreceptors : A reply to miller and bernard vision res.," *Vision Res.*, vol. 24, pp. 281–282, 1984.

[26] R. E. Kronauer and Y. Y. Zeevi, "Reorganization and diversification of signals in vision," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, pp. 91–101, Jan./Feb. 1985.

[27] J. J. Clark, M. R. Palmer, and P. D. Lawrence, "A transformation method for the reconstruction of functions from nonuniformly spaced samples," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. 33, pp. 1151–1165, Oct. 1985.

[28] Y. Zeevi and E. Shlomot, "Nonuniform sampling and antialiasing in image representation," *IEEE Trans. on Signal Processing*, vol. 41, pp. 1223–1236, March 1993.

[29] M. S. Pattichis and A. C. Bovik, "AM-FM expansions for images," in *Proc. European Signal Processing Conf.*, (Trieste, Italy), Sep. 1996.

[30] S. Qian and D. Chen, *Joint Time-Frequency Analysis: Methods and Applications.* Upper Saddle River, NJ: Prentice-Hall, 1996.

[31] A. C. Bovik, N. Gopal, T. Emmoth, and A. Restrepo, "Localized measurement of emergent image frequencies by Gabor wavelets," *IEEE. Trans. Info. Theory*, vol. IT-38, pp. 691–712, March 1992.

[32] M. Porat and Y. Y. Zeevi, "The generalized gabor scheme of image representation in biological and machine vision," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, pp. 452–468, July 1988.

[33] R. W. Hamming, *Digital Filters.* New Jersey: Prentice-Hall, Inc., 1989.

[34] E. Chang and C. K. Yap, "A wavelet approach to foveating images," in *ACM Symposium on Computational Geometry*, vol. 13, pp. 397–399, 1997.

[35] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Trans. Image Processing*, vol. 10, pp. 1397–1410, Oct. 2001.

[36] P. L. Silsbee, A. C. Bovik, and D. Chen, "Visual pattern image sequence coding," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 3, pp. 291–301, Aug. 1993.

[37] S. Daly, K. Matthews, and J. Ribas-Corbera, "Visual eccentricity models in face-based video compression," in *Proc. of SPIE, (Human Vision and Electronic Imaging IV)*, vol. 3644, (San Jose), Jan. 1999.

[38] A. Eleftheriadis and A. Jacquin, "Automatic face location detection for model-assisted rate control in h.261-compatible coding of video," *Signal Processing and Image Comm.*, vol. 7, pp. 435–455, 1995.

[39] J. Hartung, A. Jacquin, J. Pawlyk, J. Rosenberg, H. Okada, and P. E. Crouch, "Object-oriented H.263 compatible video coding platform for conferencing applications," *IEEE J. Selected Areas Commun.*, vol. 16, Jan. 1998.

[40] W. Rabiner and A. Jacquin, "Motion-adaptive modelling of scene content for very low bit rate model-assisted coding of video," *J. Visual Comm. Image Represent.*, vol. 8, pp. 250–262, Sep. 1997.

[41] S. Lee, A. C. Bovik, and B. L. Evans, "Efficient implementation of foveation filtering," in *Proc. Texas Instru-*

ments DSP Educator's Conference, (Houston, TX), Aug. 1999.

[42] J. R. Jain and A. K. Jain, "Displacement measurement and its application in interframe image coding," *IEEE Trans. Commun.*, vol. COM-29, pp. 1799–1808, Dec. 1981.

[43] M. Bierling, "Displacement estimation by hierarchical blockmatching," in *Proc. SPIE, Visual Communications and Image Processing'88.*, vol. 1001, pp. 942–951, 1988.

[44] T. Hanamura, W. Kameyama, and H. Tominaga, "Hierarchical coding scheme of video signal with scalability and compatibility," *Signal Processing and Image Comm.*, pp. 159–184, Feb. 1993.

[45] G. Gupta and C. Chakrabarti, "Architectures for hierarchical and other block matching algorithms," *IEEE Trans. Circuits Syst. Video Tech.*, pp. 477–489, Dec. 1995.

[46] A. Basu and K. J. Wiebe, "Videoconferencing using spatially varying sensing with multiple and moving fovea," in *Proc. the 12th IAPR Int. Conf. on Pattern Recognition*, (Jerusalem, Israel), pp. 30–34, Oct. 1994.

[47] "Draft text of recommendation H.263 version 2 ("h.263+") for decision," tech. rep., ITU Telecom. Standardization Section of ITU, Sep. 1997.

[48] S. Lee and A. C. Bovik, "Foveated video demonstration," in *http://pineapple.ece.utexas.edu /class/Video/demo.html*, 1999.

TABLE II

AVERAGE NUMBER OF MULTIPLICATIONS PER PIXEL FOR FOVEATION FILTERING.

	separable even symm.			circularly symm.		
	τ			τ		
	0.15	0.1	0.05	0.15	0.1	0.05
σ=0.38	7.41	10.15	20.93	8.40	16.17	69.03
σ=0.57	7.21	9.84	20.08	7.96	15.24	63.94

TABLE III

THE AVERAGE OPERATION NUMBER FOR EACH MACROBLOCK WHEN THE FOVEATION POINT DISTRIBUTION $\sigma = 0.57$. THE NOTATION LV. (%) MEANS THE PERCENTAGE OF MACROBLOCKS WITH LEVEL 1 (LV 1) TO LEVEL 5+ (LV 5+ INCLUDES THE LEVELS LARGER THAN 4)

		LV1	LV2	LV3	LV4	LV5+
HBMA 1	LV. (%)	13.2	15.5	18.0	17.6	35.7
	op. #	32496	2583	865	644	1094
	op. (%)	86.2	6.9	2.3	1.7	2.9
HBMA 2	LV. (%)	0	13.2	15.5	18.0	17.6
	op. #	0	2205	745	660	1632
	op. (%)	0.0	42.0	14.2	12.6	32.2

TABLE IV

THE AVERAGE VALUE OF THE MAD AND THE FMAD PER MACROBLOCK FOR THE REGULAR/FOVEATED "MOBILE" SEQUENCE

	reg. video		fov. video		fov. video	
			sep. symm.		cir. symm.	
	MAD	FMAD	MAD	FMAD	MAD	FMAD
FS	2087.0	243.7	918.1	156.9	867.3	148.8
HBMA 1	2219.7	248.2	921.7	157.2	870.2	148.9
HBMA 2	2310.1	253.6	937.7	158.2	883.2	149.8

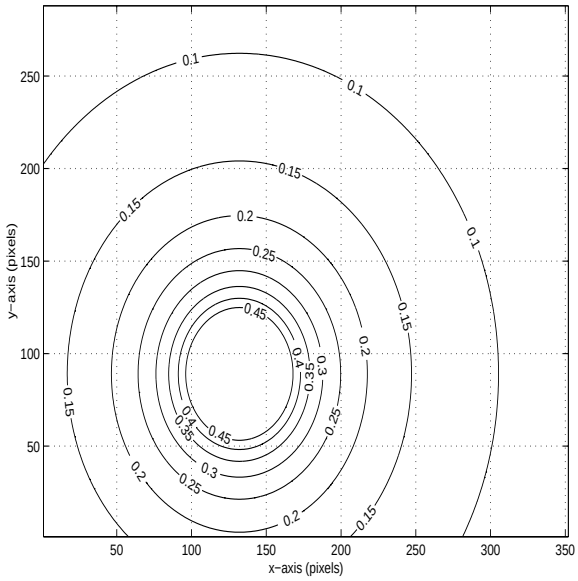
TABLE V

PREDICTION ERROR AND PICTURE QUALITY OF MOTION COMPENSATED FRAME WHEN FOVEATION POINTS ARE CHANGED

	point #	MAD	FMAD	PSNR	FPSNR
normal	$3 \rightarrow 1$	667.00	299.40	32.12	28.96
proposed	$3 \rightarrow 1$	440.78	131.74	35.69	34.56
normal	$1 \rightarrow 2$	982.46	280.83	29.65	28.01
proposed	$1 \rightarrow 2$	813.37	273.26	31.60	28.28

(a) Original *Mobile* image

(b) Foveated *Mobile* image



(c) Local bandwidth

(d) Foveated *Mobile* image over uniform domain

Fig. 1. Original vs. foveated images

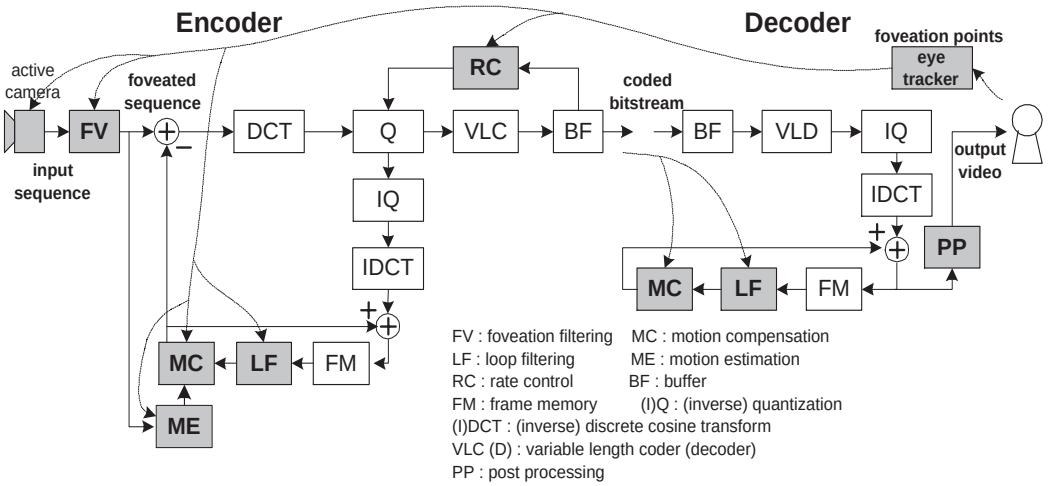
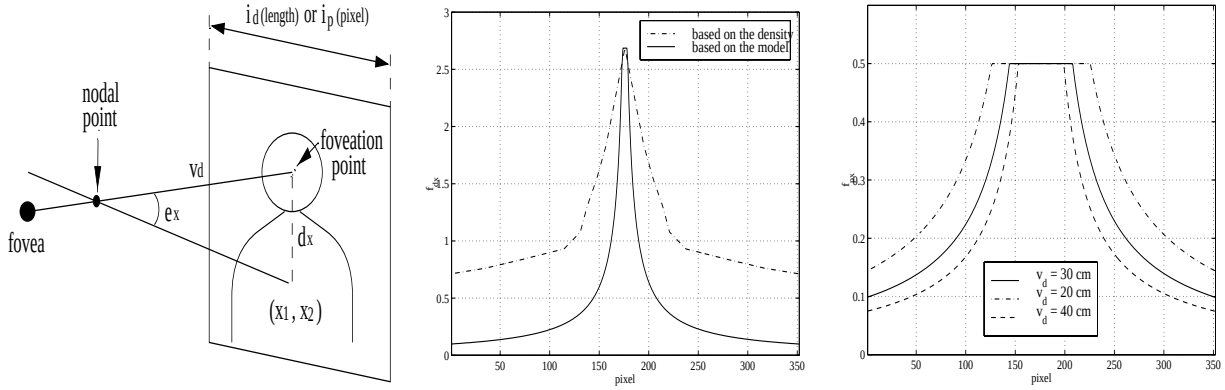


Fig. 2. An end-to-end foveated visual communication system and the foveation point information flow



(a) (b) (c)

Fig. 3. Human visual modeling : (a) Parameter definition in the viewing geometry (left), (b) Human visual detectable frequencies derived based on the photoreceptor density and the human visual model (middle), (c) Local bandwidth (in cycles/pixel) which is also used as cutoff frequencies in the filter bank (right)

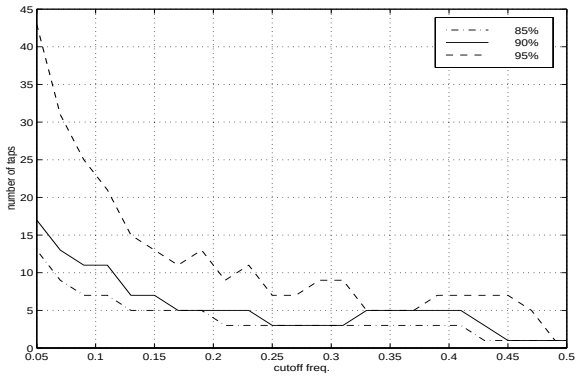


Fig. 4. The number of filter taps according to cutoff frequency

(a) Original *Lena* image

(b) Foveated *Lena* image using circularly symmetric filters with $N = 31$

(c) Foveated *Lena* image using separable even symmetric filters with adaptive N_n and $\tau = 0.1$

(d) Foveated *Lena* image using circularly symmetric filters with adaptive N_n and $\tau = 0.1$

Fig. 5. Original "Lena" image v.s. foveated "Lena" images

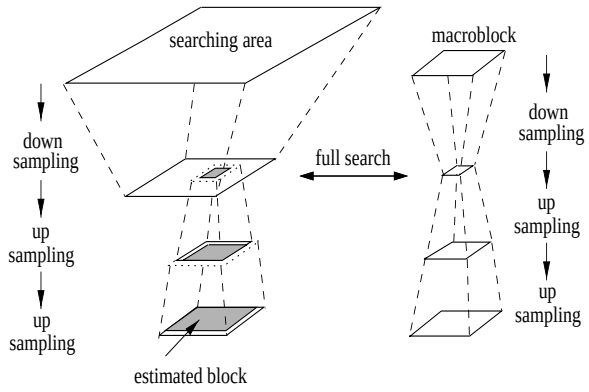
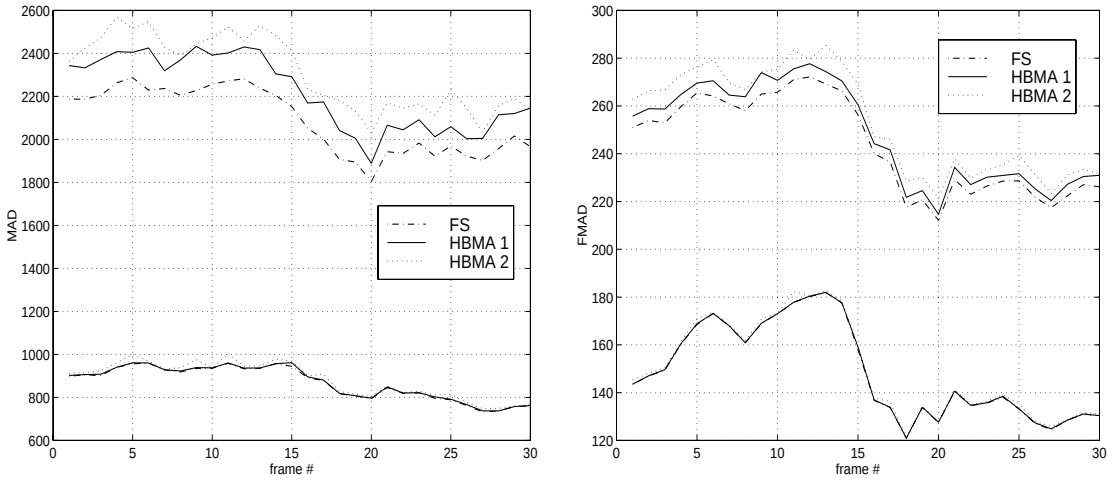
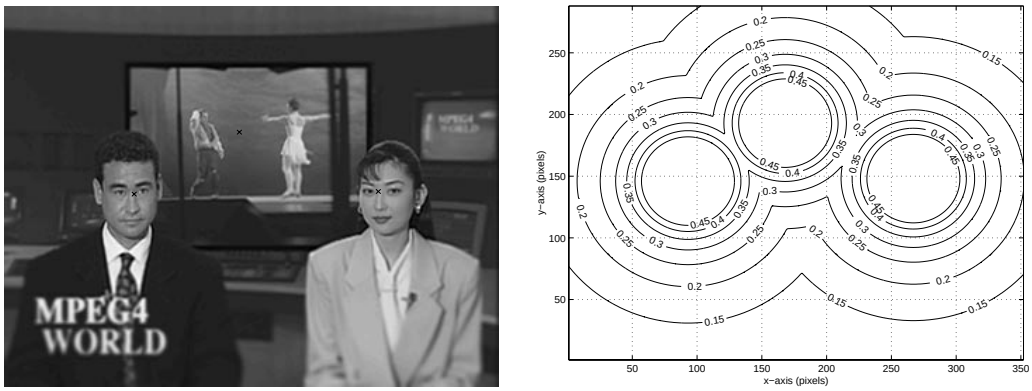


Fig. 6. Hierachical block diagram for the foveated video



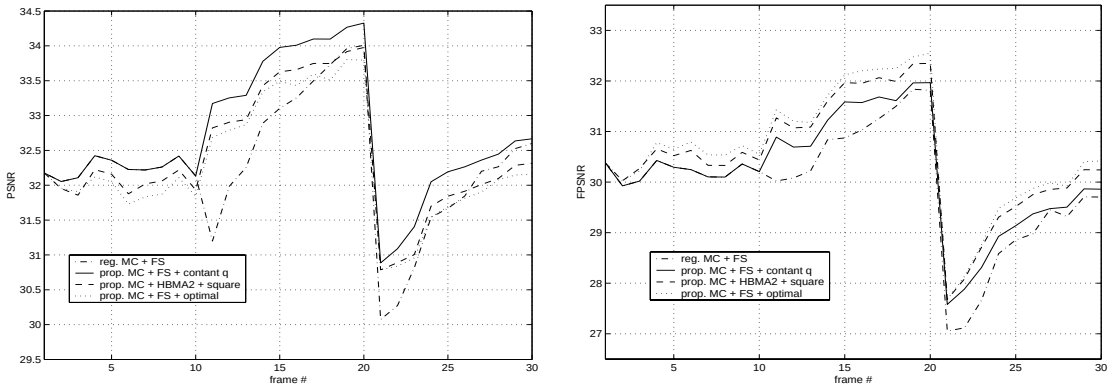
(a) MAD (b) FMAD

Fig. 7. The MAD (a) and the FMAD (b) of motion compensated frames according to "Mobile" frames. The 3 upper lines are for the regular video and the 3 lower lines are for the foveated video.



(a) Foveated image *News* (b) Local bandwidth

Fig. 8. Foveated "News" images(352x288) with 3 foveation points (93,143), (267,141), (168,96) and local bandwidth



(a) PSNR v.s. frame # (b) FPSNR v.s. frame #

Fig. 9. Foveation points are changed from $\{p_1, p_2, p_3\}$ to $\{p_1\}$ in the 11^{th} frame and to $\{p_3\}$ in the 21^{th} frame

(a) I frame of regular image, 277 $Kbits$, PSNR = 29.9 dB, FPSNR = 28.9 dB, rate control : *const q*, QP = 25

(b) I frame of regular image, 278 $Kbits$, PSNR = 29.5 dB, FPSNR = 29.9 dB, rate control : *square*, $Q_m = 14$

(c) I frame of foveated image, 271 $Kbits$, PSNR = 33.0 dB, FPSNR = 32.6 dB, rate control : *square*, $Q_m = 12$

(d) I frame of foveated image, 276 $Kbits$, PSNR = 33.5 dB, FPSNR = 32.3 dB, rate control : *linear*, $Q_m = 15$

Fig. 10. The reconstructed "Lena" images according to the rate control methods

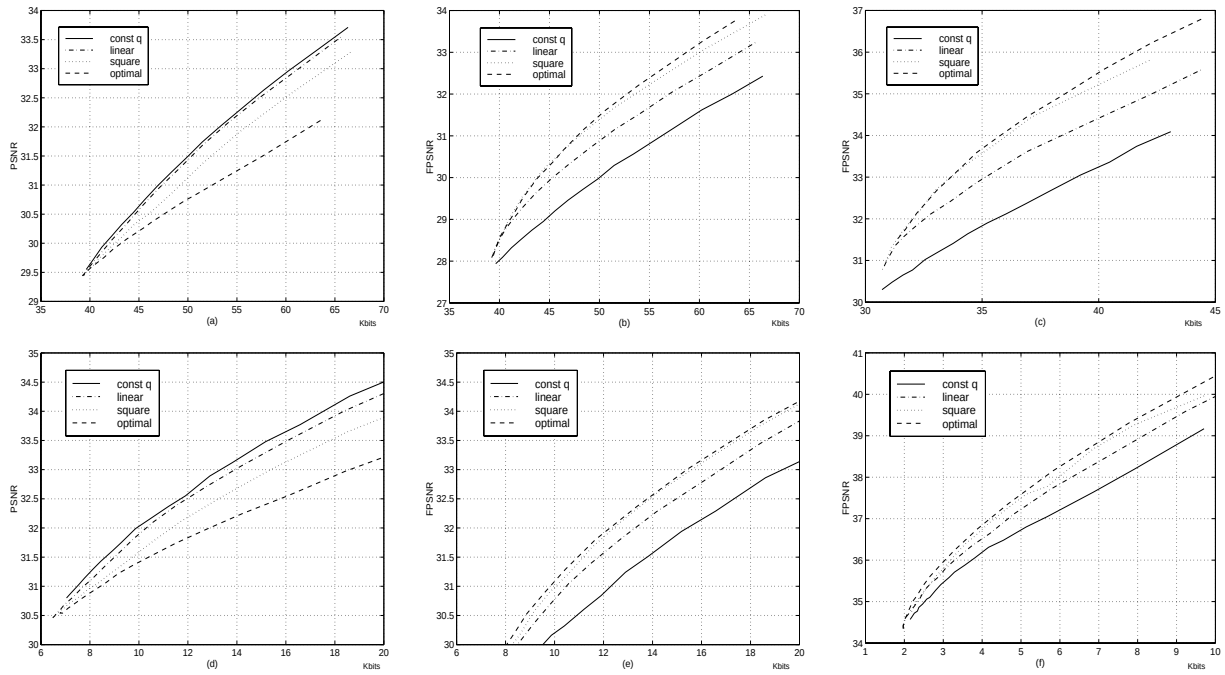


Fig. 11. The PSNR/FPSNR for video sequences : (a) PSNR : "Mobile" I frame, (b) FPSNR : "Mobile" I frame, (c) FPSNR : "News" I frame, (d) PSNR : "Mobile" P frame, (e) FPSNR : "Mobile" P frame, (f) FPSNR : "News" P frame

Fig. 12. The postprocessing version of reconstructed images of Fig. 10 (a) in the left image, and of Fig. 10 (c) in the right image