Foveation Scalable Video Coding with Automatic Fixation Selection

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

Abstract—**Image and video coding is an optimization problem. A successful image and video coding algorithm delivers a good tradeoff between visual quality and other coding performance measures, such as compression, complexity, scalability, robustness, and security. In this paper, we follow two recent trends in image and video coding research. One is to incorporate human visual system (HVS) models to improve the current state-of-the-art of image and video coding algorithms by better exploiting the properties of the intended receiver. The other is to design rate scalable image and video codecs, which allow the extraction of coded visual information at continuously varying bit rates from a single compressed bitstream.**

Specifically, we propose a foveation scalable video coding (FSVC) algorithm which supplies good quality-compression performance as well as effective rate scalability. The key idea is to organize the encoded bitstream to provide the best decoded video at an arbitrary bit rate in terms of foveated visual quality measurement. A foveation-based HVS model plays an important role in the algorithm. The algorithm is adaptable to different applications, such as knowledge-based video coding and video communications over time-varying, multi-user and interactive networks.

Keywords—**video coding, rate scalable coding, human visual system, foveation, image and video quality, wavelet**

I. Introduction

It has been envisioned that network visual services, such as network video broadcasting, video-on-demand, video-conferencing and telemedicine, will become ubiquitous in the twenty-first century. As a result, network visual communication has become an active research area in recent years. One of the most challenging problems for the implementation of a video communication system is that the available bandwidth of the networks is usually insufficient for the delivery of the voluminous amount of the video data. In order to solve this problem, considerable effort has been applied in the last three decades for the development of video compression techniques. These efforts have resulted in the video coding standards such as H.261 [1], H.263 [2], MPEG-1 [2], [3], MPEG-2 [2], [3], and MPEG-4 [2], [4].

Designing a video coding and communication system is a complicated task. The first issue that needs to be considered is the quality-compression performance, which aims to provide the best quality decoded video with the minimal number of bits. Depending on the application, there are many other issues related to the goodness of the video codecs. For example, low computational complexity is usually required for real-time applications. In many cases, parallelizability is a desired feature to improve speed. Satisfying a low memory requirement is desirable in many applications to achieve easy buffering and easy embedded implementations on digital signal processors. Several communication and networking issues are also relevant, such as scalability, robustness, security and interactivity. Although the video coding standards exhibit acceptable quality-compression performance in many visual communication applications, further improvements are desired and more features need to be added, especially for some specific applications. Recently, two interesting research trends have emerged that are very promising and may lead to significantly improved video codecs in comparison with the current standards.

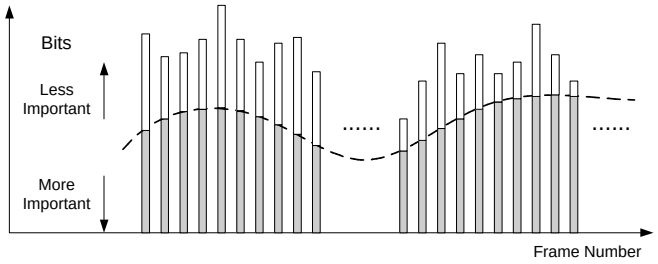

Fig. 1. Bitstream scaling in rate scalable video communications. Each bar represents the bitstream for one frame in the video sequence.

The first trend is to incorporate Human Visual System (HVS) models into the coding system. Presently, the objective quality measure Peak Signal-to-Noise Ratio (PSNR) is widely employed to evaluate video quality. However, it is well accepted that perceived video quality does not correlate well with PSNR. HVS characteristics must be considered to provide better visual quality measurements. In the literature, many HVS-based algorithms have been proposed for this purpose [5]–[11]. Although the current understanding of the HVS still is insufficient to provide a precise, generic and robust algorithm to measure perceived video quality in all circumstances, it is believed that an appropriate HVS model that takes advantage of some well-understood HVS features can significantly help to improve the current state-of-the-art of video coding algorithms.

The second research trend is to develop continuously rate scalable coding algorithms [12]–[20], which allow the extraction of coded visual information at continuously varying bit rates from a single compressed bitstream. An ex-

Z. Wang was with Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin, Austin, TX 78712. He is now with Laboratory for Computational Vision (LCV), New York University, New York, NY 10003. L. Lu is with Multimedia Technologies, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598. A. C. Bovik is with Laboratory for Image and Video Engineering (LIVE), The University of Texas at Austin, Austin, TX 78712. E-mail: zhouwang@ieee.org, lul@us.ibm.com, bovik@ece.utexas.edu.
This research was supported in part by IBM Corp., Texas Instruments, Inc., and by State of Texas Advanced Technology Program.

ample is shown in Fig. 1, 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 [2],[3],[21], video transcoding [22],[23], 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.

In this paper, we propose a new video coding approach called Foveation Scalable Video Coding (FSVC), which stands at the intersection of the two promising research trends. Specifically, wavelet-based embedded bitplane coding techniques are used for rate scalable coding. Further, we exploit the foveation feature of the HVS, which refers to the fact that the HVS is a highly space-variant system, where the spatial resolution is highest at the point of fixation (foveation point) and decreases dramatically with increasing eccentricity. By taking advantage of the this effect, considerable high frequency information redundancy can be removed from the peripheral regions without significant loss of the reconstructed image and video quality. The foveation factor has been employed in previous work to improve image and video coding efficiency [24]–[36]. Foveated image and video coding is closely related to Region-of-Interest (ROI) image and video coding (e.g., [37]–[40]). "If we define the area(s) around the point(s) of fixation as the region of interest, then foveation-based image processing can be viewed as a special case of ROI image processing" [41]. The major difference with respect to traditional ROI processing is that the "interest" is continuously space-variant and conforms with HVS characteristics. Most of the foveation algorithms used a fixed foveation model. These methods lack the flexibility to adapt to different foveation depths and are not convenient to be implemented in a rate scalable manner. Chang *et al.* made one of the first attempts to develop a wavelet-based scalable foveated image compression and progressive transmission system [42]–[44]. However, human visual characteristics were not considered in depth, and no efficient coding algorithms were implemented to provide a quality-compression performance comparable to other state-of-the-art of image coding techniques.

In [20], Wang and Bovik proposed a scalable foveated wavelet image coding algorithm termed Embedded Foveation Image Coding (EFIC), which naturally combines foveation filtering with foveated image compression and provides very good coding performance in terms of foveated visual quality measurement. This paper attempts to extend the work in [20] for video coding. There are two major purposes. The first is to establish a prototype for rate scalable foveated video coding. The prototype must be very flexible such that different foveation point selection schemes can be applied to a single framework. The second purpose is to implement the prototype in a specific application environment, where FSVC is combined with an automated foveation point selection scheme and an adaptive frame prediction algorithm.

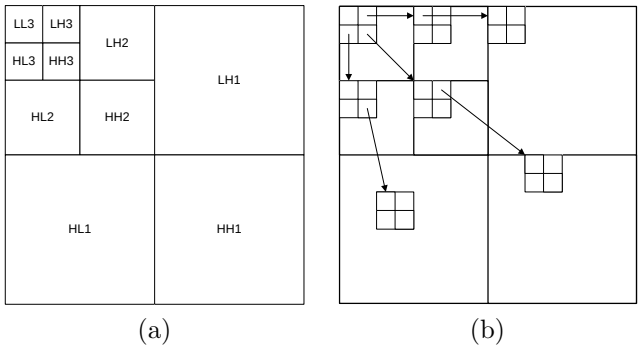

Fig. 2. (a) 2-D DWT decomposition structure; (b) Spatial orientation tree in SPIHT algorithm.

In Section II, we describe briefly the basic methods of wavelet-based embedded bitplane coding and introduce the general framework of our FSVC system. Section III develops the foveation-based HVS model. More details about the implementation of the FSVC algorithm are given in Section IV. Finally, Section V makes some concluding remarks and provides further discussions.

II. Basic Methods and General Framework

A. Wavelet-Based Image and Video Compression and Embedded Encoding

Recently, wavelet-based methods have achieved great success in still image coding [12]–[14], [20], [45]. The success is due to the energy compaction feature of the Discrete Wavelet Transforms (DWTs) and the efficient organization, quantization, and encoding of the wavelet coefficients. Wavelet-based methods have also been applied to compress video [15]–[19], [46]. Readers can refer to [47], [45], [48] and [49] for more introductory information about wavelets, wavelet transforms, and how wavelet transforms are used for image and video compression.

A class of embedded bitplane coding algorithms has recently attracted great attention. The most well-known algorithms include Shapiro's Embedded Zerotree Wavelet (EZW) algorithm [18], and Said and Pearlman's Set Partitioning Into Hierarchical Trees (SPIHT) algorithm [19], which is a refined implementation of the EZW idea. The main objective of embedded wavelet coding is to order the output bistream, such that the bits with greater contribu-

tion to the Mean Squared Error (MSE) between the original and the compression images are encoded and transmitted first. It has been observed that the wavelet coefficients have structural similarity across the wavelet subbands in the same spatial orientation. The zero tree structure in EZW and the spatial orientation tree structure in SPIHT capture this structural similarity very effectively. For a 2-D DWT decomposition shown in Fig. 2(a), the spatial orientation tree used by SPIHT is given in 2(b). In the EZW and SPIHT encoders, 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 is usually used to further compress the output bitstream.

In HVS-based wavelet image coding algorithms, the wavelet coefficients are usually weighted according to visual importance before the encoding procedures [10], [20], [50], [51]. In [20], a modified SPIHT algorithm is designed to improve the coding efficiency for weighted wavelet coefficients.

B. Foveation Point(s) Setup

The basic idea of FSVC is to order the output bistream, such that the information associated with the foveated areas have higher priorities to be encoded earlier. Before the encoding process, however, it is necessary to select the foveation point(s). The best way of foveation point(s) selection is highly application dependant. We attempt to have a flexible design in FSVC, so that it can be used in various cases.

First, we allow FSVC to select multiple foveation points. The reason is multifold:

1) The usual pattern of human fixation is that the fixation point moves slightly within a small area around the center point of interest [52];

2) There may be multiple human observers watching the image at the same time;

3) There may exist multiple points and/or regions in the image that have high probability to attract a human observer's attention.

4) Certain foveation points can be put at areas where the human eyes are very sensitive to distortions. This is actually an extension of the foveation model, so that other HVS features can be included into the same framework.

Second, we limit the search space of the foveation points. Theoretically, any pixel in the observed picture could be visually foveated. In practice, however, testing all the possible pixels require very high computation power. Also, encoding the locations of the foveation points will consume many bits, leading to significant overhead in the encoded bitstream. Further, since small shifts of the foveation points will not result in significant difference in visual quality and system encoding performance, it is not worth spending too many bits and computation power to generate

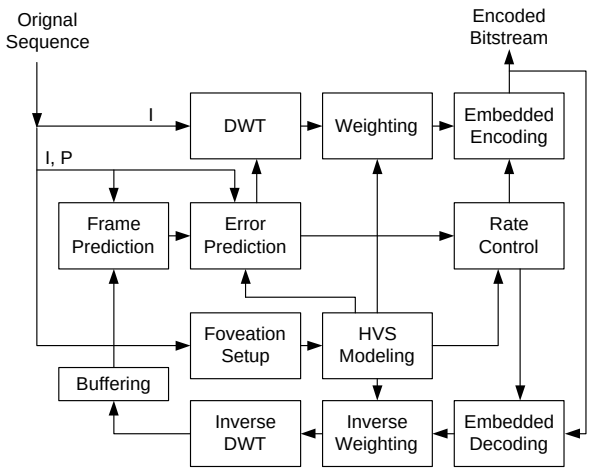

Fig. 3. General framework of the FSVC encoding system.

and encode the foveation point locations and to calculate the foveated HVS model. The FSVC system first divides the picture being encoded into blocks with a size of 16×16, and the candidate foveation points are limited to the centers of these blocks. By using this strategy, computation is considerably reduced and only one bit for each block is needed to encode the foveation point selection information. Using this method, a binary map with a size of $\frac{1}{16 \times 16}$ of the original image will be generated. This map can be further compressed with an entropy coding technique such as the arithmetic coding algorithm [53].

C. Framework of the Encoding System

Similar to many other video coding methods, FSVC first divides the 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 general framework for the encoding of I frames and P frames is given in Fig. 3.

The encoding of the I frame is the same as the EFIC algorithm [20] developed for still image coding. Firstly, apply the DWT and obtain the wavelet coefficients. Secondly, the foveation point selection scheme is applied and the HVS model is calculated to determine the visual importance of the wavelet coefficients. The importance value of each wavelet coefficient is then used to weight the wavelet coefficient. Finally, the modified SPIHT algorithm [20] is employed to generate the embedded bitstream.

The encoding of the P frames is more complicated. The idea of using P frames in video coding is to exploit temporal redundancy between adjacent frames in the video sequence. Prediction of the current frame from its previous frame is the key technique to make use of temporal redundancy. Motion Estimation (ME) and Motion Compensation (MC) techniques have been successfully used for this purpose. The main difference between our FSVC algorithm and other video coding algorithms is that it uses two instead of one version of the previous frames. One is the original previous frame. 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 foveation-based HVS model, which will be discussed in detail in Section III. The DWT is applied to the prediction error frame, and the resulting coefficients are weighted and coded with the embedded encoding algorithm.

The HVS modelling techniques are different for I frames and P frames. This will be discussed in Section III. During the encoding process, a rate control algorithm is used to allocate bits to each frame. The allocation is determined by the available bandwidth, user requirements, the HVS modelling results and the frame prediction error.

III. Foveation Based HVS Model

A. Foveated Resolution and Sensitivity Model

Psychological experiments have been conducted to measure the contrast sensitivity as a function of retinal eccentricity [32], [54]–[56]. In [32], a model that fits the experimental data was given by

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

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 [32] are $\alpha = 0.106$, $e_2 = 2.3$, and $CT_0 = 1/64$. The contrast sensitivity is defined as the reciprocal of the contrast threshold:

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

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

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

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. 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 $d(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}^f\|_2 = \sqrt{(x_1 - x_1^f)^2 + (x_2 - x_2^f)^2}$ (measured in pixels). The eccentricity is then calculated as

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

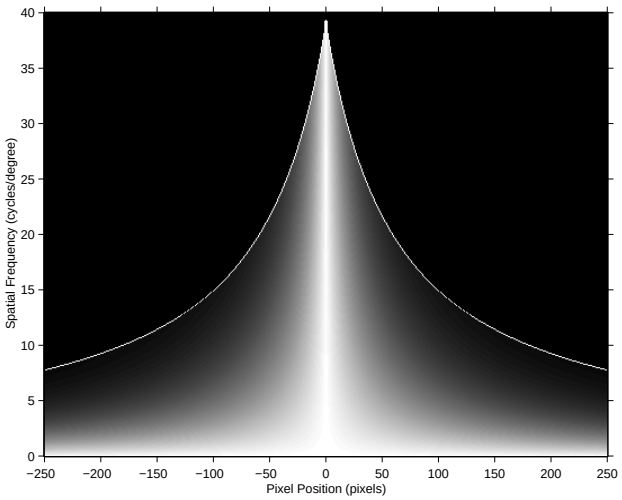

Fig. 4. Normalized contrast sensitivity (Brightness indicates the strength of contrast sensitivity) for $N = 512$ and $v = 3$. The white curves show the cutoff frequency.

With (4), we can convert the foveated contrast sensitivity and cutoff frequency models into the image pixel domain. In Fig. 4, we show the normalized 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:

$$r \approx \frac{\pi N v}{180} \left(\frac{pixels}{degree}\right) . \quad (5)$$

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:

$$f_d(v) = \frac{r}{2} \approx \frac{\pi N v}{360} \left(\frac{cycles}{degree}\right) . \quad (6)$$

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

$$f_m(v,\mathbf{x}) = min(f_c(e(v,\mathbf{x})), f_d(v)) . \quad (7)$$

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

$$S_f(v,f,\mathbf{x}) = \begin{cases} \frac{CS(f,e(v,\mathbf{x}))}{CS(f,0)} & if \quad f \le f_m(v,\mathbf{x}) \\ 0 & otherwise \end{cases} . \quad (8)$$

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

B. Spatial Domain Foveated Weighting Model

In the FSVC system, two foveated weighting models are developed, one in the spatial domain and the other in the

DWT domain. The spatial domain weighting model is employed by the adaptive frame prediction algorithm to adjust the combination from the original and the decoded motion-compensated frames, and the wavelet domain weighting model is used to determine the importance of the wavelet coefficients and help ordering the output bitstream.

The spatial domain weighting model is obtained by normalizing the cutoff frequency model defined in (7):

$$W_s(v,\mathbf{x}) = \left[\frac{f_m(v,\mathbf{x})}{f_m(v,\mathbf{x}^f)}\right]^{\gamma} , \tag{9}$$

where γ is a parameter used to control the shape of the weighting model. For a fixed viewing distance v_0, this weighting model can be written as $W_s(\mathbf{x}) = W_s(v_0,\mathbf{x})$. This model can easily adapt to multiple foveation points. Suppose that there are K foveation points $\mathbf{x}_1^f$, $\mathbf{x}_2^f$, $\cdots$, $\mathbf{x}_K^f$ in the image. For each of the points, we can calculate the weighting model individually and have $W_s^i(\mathbf{x})$ for $i = 1, 2, \cdots, K$. In the worst case, the human observer would fixate at the foveation point which is the closest with respect to $\mathbf{x}$. This results in the maximum value of $W_s^i(\mathbf{x})$ for all i. Therefore, the overall weighting value for $\mathbf{x}$ is given by:

$$W_s(\mathbf{x}) = \max_{i\in\{1,\cdots,K\}} W_s^i(\mathbf{x}) . \tag{10}$$

In practice, it is not necessary to compute each of the $W_s^i(\mathbf{x})$. Because for a given pixel $\mathbf{x}$, the foveation point that is closest to it must generate the maximum weight, hence we have

$$W_s(\mathbf{x}) = W_s^j(\mathbf{x}), \quad j \in \arg\min_{i\in\{1,\cdots,K\}} \left\{ \left\| \mathbf{x} - \mathbf{x}_i^f \right\|_2 \right\} . \tag{11}$$

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

C. Wavelet Domain Foveated Weighting Model

The wavelet coefficients at different subbands and locations supply information of variable perceptual importance to the HVS. In [10], 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 [10]:

$$\log Y = \log a + k(\log f - \log g_\theta f_0)^2 \tag{12}$$

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 [10]: $f = r2^{-\lambda}$. The constant parameters in (12) 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:

$$T_{\lambda,\theta} = \frac{Y_{\lambda,\theta}}{A_{\lambda,\theta}} = \frac{a10^{k\left[\log(r)-\log(2^\lambda g_\theta f_0)\right]^2}}{A_{\lambda,\theta}} , \tag{13}$$

where $A_{\lambda,\theta}$ is the basis function amplitude given in [10]. We define the error sensitivity in subband (λ,θ) as:

$$S_w(\lambda,\theta) = \frac{1}{T_{\lambda,\theta}} . \tag{14}$$

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

$$d_{\lambda,\theta}(\mathbf{x}) = 2^\lambda \left\| \mathbf{x} - \mathbf{x}_{\lambda,\theta}^f \right\|_2 \qquad for \quad \mathbf{x} \in \mathbf{B}_{\lambda,\theta} , \tag{15}$$

where $\mathbf{x}_{\lambda,\theta}^f$ is the corresponding foveation point in subband (λ,θ). With the equivalent distance, and also considering (8), we have

$$S_f(v,f,\mathbf{x}) = S_f(v, r2^{-\lambda}, d_{\lambda,\theta}(\mathbf{x})) \qquad for \quad \mathbf{x} \in \mathbf{B}_{\lambda,\theta} . \tag{16}$$

Combining (14) and (16), a wavelet domain foveation-based visual sensitivity model is achieved:

$$S(v,\mathbf{x}) = [S_w(\lambda,\theta)]^{\beta_1} \cdot \left[S_f(v, r2^{-\lambda}, d_{\lambda,\theta}(\mathbf{x}))\right]^{\beta_2} \quad \mathbf{x} \in \mathbf{B}_{\lambda,\theta} , \tag{17}$$

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:

$$W_w(\mathbf{x}) = \int_{0^+}^{\infty} p(v) S(v,\mathbf{x})\, dv , \tag{18}$$

where $p(v)$ is the probability density distribution of the viewing distance v [20]. Fig. 5 shows the final importance weighting mask in the DWT domain. Similar to the spatial domain model, for the case of multiple foveation points, the overall weighting value is obtained by:

$$W_w(\mathbf{x}) = W_w^j(\mathbf{x}), \quad j \in \arg\min_{i\in\{1,\cdots,K\}} \left\{ \left\| \mathbf{x} - \mathbf{x}_{i,\lambda,\theta}^f \right\|_2 \right\} . \tag{19}$$

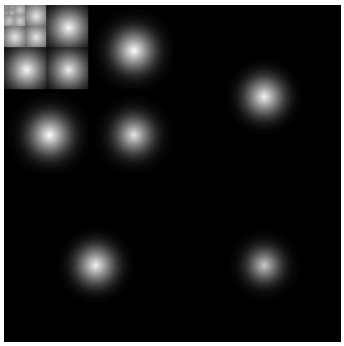

Fig. 5. Wavelet domain importance weighting mask of a single foveation point. Brightness indicates the importance of the wavelet coefficients (Brightness logarithmically enhanced for display purpose).

(a) (b)

Fig. 6. An I frame (a) and a P frame (b) in the "News" video sequence.

IV. Implementation of FSVC

The general framework introduced in Section II is flexible and can adapt to different application environments. As an example, our current implementation of FSVC focuses on developing an automated foveation setup approach for video sequences with human faces. Furthermore, an adaptive algorithm is proposed for the prediction of the current frame from motion compensated previous frames.

A. Determination of Foveation Points

Human faces are probably the most frequently focused regions by human observers. A face-foveated video coding algorithm will be very useful to effectively enhance the visual quality in many specific video communication environments such as videoconferencing. The face detection algorithm used in our FSVC implementation is similar to that in [57]. It consists of three steps.

The first step is to identify the possible face regions by the skin color information [58]. The entire YCrCb color space is divided into a skin color subspace and a non-skin color subspace. Each point in the picture can then be assigned to either of the two subspaces.

In Step 2, we detect human faces in those skin-color regions by a technique called binary template matching [57].

In the last step, we verify every detected face and remove falsely detected faces. The verification is based on the observation that human face areas usually have a certain amount of high frequency content [57] because of the existing of discontinuities at eyes, nose and mouth. For each detected face region, we calculate the variance of the pixels in it. Only the regions with variances larger than a threshold value are finally verified as face regions.

The methods to select foveation points for I frames and P frames are different. For I frames, we first detect face areas as regions of interest and put foveation points inside those regions. The face detection algorithm described above is very efficient but does not provide precise boundaries of the face areas. Since small shifts of foveation points do not have significant effects on visual quality, this kind of rough face detection is enough for the FSVC system to work properly. An example is given in Fig. 6 and Fig. 7, where one I frame extracted from the "News" sequence is shown in Fig. 6(a). The selected foveation points of this frame is given in Fig. 7(a).

For the foveation point selection of P frames, a different strategy is used because of two reasons. First, more information is available because the current P frame can be compared with the previous frame to locate the new information presented in the current frame. Second, the P frames are not encoded directly. Only the difference between the current frame and the prediction from the previous frame is of concern to us. If the prediction error of a local region is very small, then it is not necessary to put any foveation point in that region, regardless of whether the region is fixated or not. FSVC focuses on the regions in the current P frame that provide us with new information from its previous frame. Usually, the prediction errors in those regions are larger than other regions. Therefore, FSVC mainly selects foveation points in those regions with prediction errors larger than a threshold value. For example, for the P frame shown in Fig. 6(b), which follows the I frame given in Fig. 6(a), the error thresholding-based method selects the foveation regions shown in Fig. 7(b). The drawback of this method is that the face regions will lose fixation. To solve this problem, we use an unequal error thresholding method to determine foveation regions in P frames. This is based on the fact that when human observers' attention is fixating on human faces, even very small changes in these areas are very likely to be noticed. Therefore, we use a much smaller prediction error threshold value to capture the changes occurring in the face regions. Using the unequal error thresholding based method, the foveation region selection result for Fig. 6(b) is shown in Fig. 7(c). Compared with Fig. 7(b), some foveation points in the face regions are added.

B. Adaptive Frame Prediction

In fixed rate ME/MC based video coding algorithms, a common choice for frame prediction is to use the feedback decoded previous frame as the reference frame for the pre-

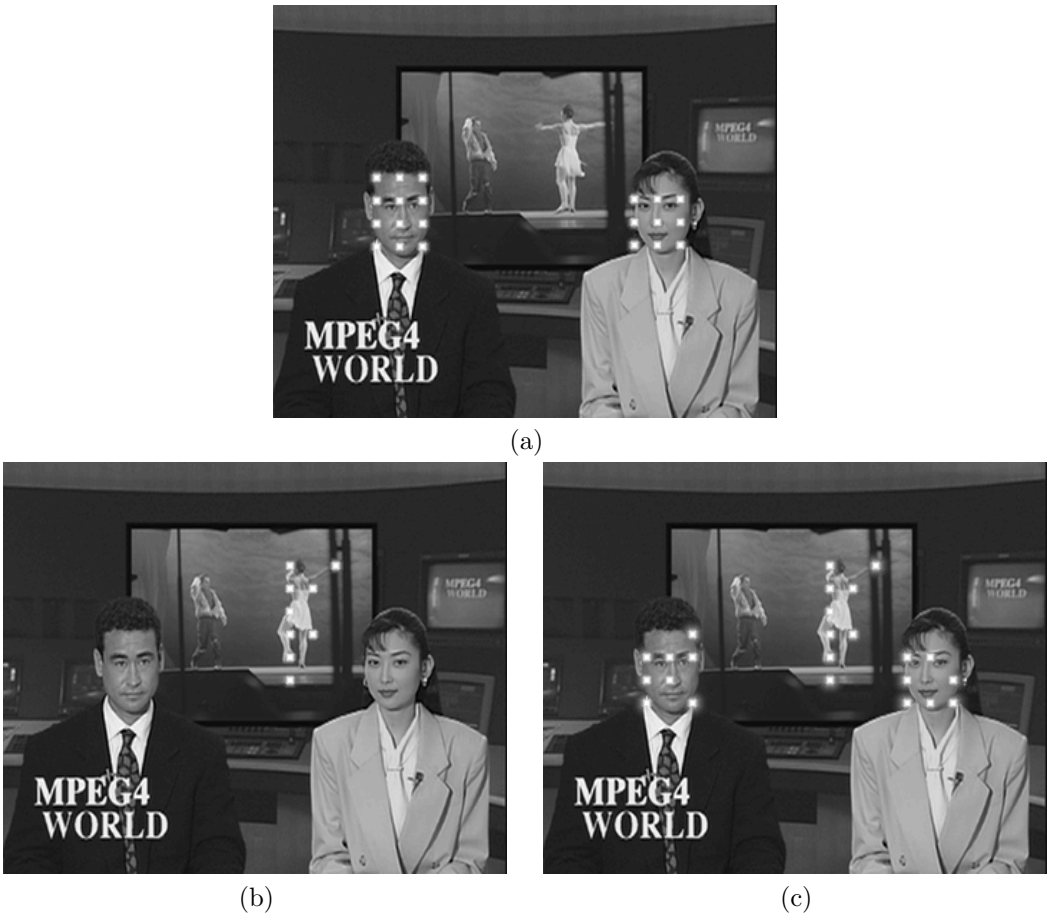

Fig. 7. Foveation point selection of the "News" video sequence. (a) I frame foveation point selection; (b) P frame foveation point selection with equal error thresholding; (c) P frame foveation point selection with unequal error thresholding.

diction of the current frame. With this choice, the prediction frames are exactly the same at the encoder and the decoder. However, this choice is infeasible for continuously rate scalable coding because the decoding bit rate is the choice of the decoder and is unavailable to the encoder. There are several solutions to this problem.

The first solution simply uses the original motion compensated frames to do the prediction. Since the original frames are not available at the decoder, the prediction frames at the encoder and the decoder sides are different, sometimes very different. The consequence is that very good frame prediction at the encoder side may produce poor prediction at the decoder side. In addition, the poor prediction error will propagate to all the following P frames in the same GOP.

The second solution is to define a low base bit rate and use the decoded and motion compensated frame at the base bit rate as the prediction. This idea has been used in [17], [59]. The advantage of this solution is that the prediction frames at the encoder and the decoder are exactly the same. Therefore, significant error propagation problems are avoided. However, if the decoding bit rate is much higher than the base bit rate, large prediction errors will occur. For example, suppose we have a texture region that does not change between frames. At an I frame, the region is encoded at a high bit rate with high quality. Since there is no change between frames, very good prediction with almost zero prediction error is expected. However, with the second prediction solution, the low base rate decoded frame (with low quality) is selected to do the prediction. This leads to poor prediction and the fine textures of the regions are actually encoded repeatedly. In conclusion, this solution results in less precise prediction and less efficient compression.

We propose a new solution to this problem, where the original motion compensated frame and the base bit rate decoded and motion compensated frame are combined to make a prediction. The combination is adaptively changed using the foveation model. The encoder and decoder sides of the new frame prediction algorithms are shown in Fig. 8 and Fig. 9, respectively.

At the encoder, there are two reference frames. One is the previous frame from the original sequence, and the other is the previous frame decoded from the base bit rate. The same motion compensation process is applied to both of them and generates two motion compensated reference frames. These two frames are combined by the spatial domain foveation weighting model. Let $W_s(\mathbf{x})$ be the normalized weight at location $\mathbf{x}$. Let $P_O(\mathbf{x})$ and $P_B(\mathbf{x})$ be the pixel values at location $\mathbf{x}$ of the motion compensated origi-

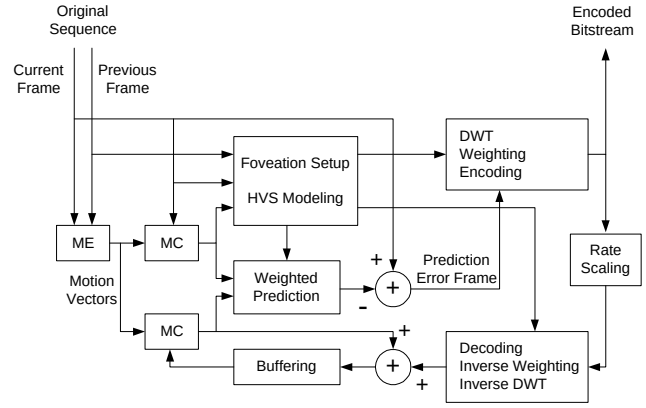

Fig. 8. Adaptive frame prediction: encoder side.

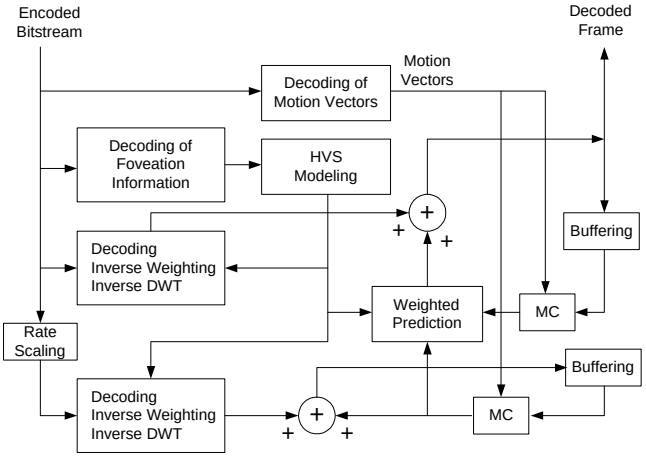

Fig. 9. Adaptive frame prediction: decoder side.

nal reference frame and base rate decoded reference frame, respectively. Then the combined encoder prediction value $P_E(\mathbf{x})$ is given by:

$$P_E(\mathbf{x}) \;=\; [1 - W_s(\mathbf{x})]\, P_O(\mathbf{x}) \;+\; W_s(\mathbf{x})\, P_B(\mathbf{x}) \;. \qquad (20)$$

At the decoder, the weighting information is decoded and calculated in exactly the same way as in the encoder. There are also two versions of the reference frames. One is the previous frame decoded from the base rate. The other is the previous frame decoded at the current decoding bit rate. Motion compensation is applied to both reference frames. Let $P_C(\mathbf{x})$ be the pixel values at location $\mathbf{x}$ of the motion compensated reference frame at the current decoding bit rate, then the combined decoder prediction value $P_D(\mathbf{x})$ is:

$$P_D(\mathbf{x}) \;=\; [1 - W_s(\mathbf{x})]\, P_C(\mathbf{x}) \;+\; W_s(\mathbf{x})\, P_B(\mathbf{x}) \;. \qquad (21)$$

The idea behind the weighting equations (20) and (21) is that for the difficult prediction regions, more weight is given to the base rate motion compensated reference frames, while for the easy prediction regions, more weight is given to the high quality motion compensated reference frames. If the prediction errors are in the mid-range, the adaptive frame prediction algorithm will provide a "fuzzy" solution according to the mid-range values of $W_s(\mathbf{x})$, which provides a trade-off between prediction from the base rate motion compensated frame and the prediction from high quality motion compensated reference frame. The frame predictions at the encoder and decoder are not exactly the same. Subtracting (21) from (20) yields

$$P_E(\mathbf{x}) \;-\; P_D(\mathbf{x}) \;=\; [1 - W_s(\mathbf{x})]\,[P_O(\mathbf{x}) \;-\; P_C(\mathbf{x})] \;. \qquad (22)$$

Since at the difficult prediction regions, the value of $W_s(\mathbf{x})$ is large, the error between $P_E(\mathbf{x})$ and $P_D(\mathbf{x})$ is very small and can be neglected. At the easy prediction regions, the values of $P_C(\mathbf{x})$ is close to $P_O(\mathbf{x})$. Therefore, the prediction difference between the encoder and the decoder is small. In this way, the error propagation is well controlled. Also note that at the easy prediction regions, the value of $W_s(\mathbf{x})$ is small and the actual prediction in (20) and (21) is mainly from $P_O(\mathbf{x})$ and $P_C(\mathbf{x})$. Since $P_O(\mathbf{x})$ and $P_C(\mathbf{x})$ are from high quality prediction frames, their prediction values are much better than the poor prediction of $P_B(\mathbf{x})$. In this way, the prediction errors are reduced.

In conclusion, by using the new frame prediction algorithm, 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.

C. Experimental Results

We test the FSVC system on CIF size (288×352), YCbCr 4:2:0 format video sequences. In order to give a quantitative measurement on how much quality gain is achieved by using the foveated techniques, it is important to employ an image quality metric designed for foveated images. Most image quality measurement methods in the literature are not appropriate because they are designed for uniform resolution images. In [20], [60], a wavelet-based foveated image quality assessment metric called Foveated Wavelet Quality Index (FWQI) was proposed by combining the wavelet domain visual sensitivity model (17) and a novel image quality indexing algorithm [61], [62]. FWQI has a dynamic range of [0, 1], where 1 represents the best quality. One distinct feature of the new quality indexing approach in comparison with the traditional image quality assessment techniques is that it considers image degradations as "structural information loss" or "structural distortions" instead of "perceived errors". More insights and discussions about the new indexing method are provided in [11], [41], [63].

Fig. 10 shows 4 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. The FSVC compression result at 200 Kbits/sec is also given in the same figure. It can be observed that the face region and the relative moving information between frames are captured very well with the automated foveation point selection algorithm.

Fig. 11 compares the compression results of the 26th frame (a P frame) of the "News" sequence to demonstrate the effectiveness of the foveation method against non-foveation method and the adaptive frame prediction scheme against traditional frame prediction schemes. The

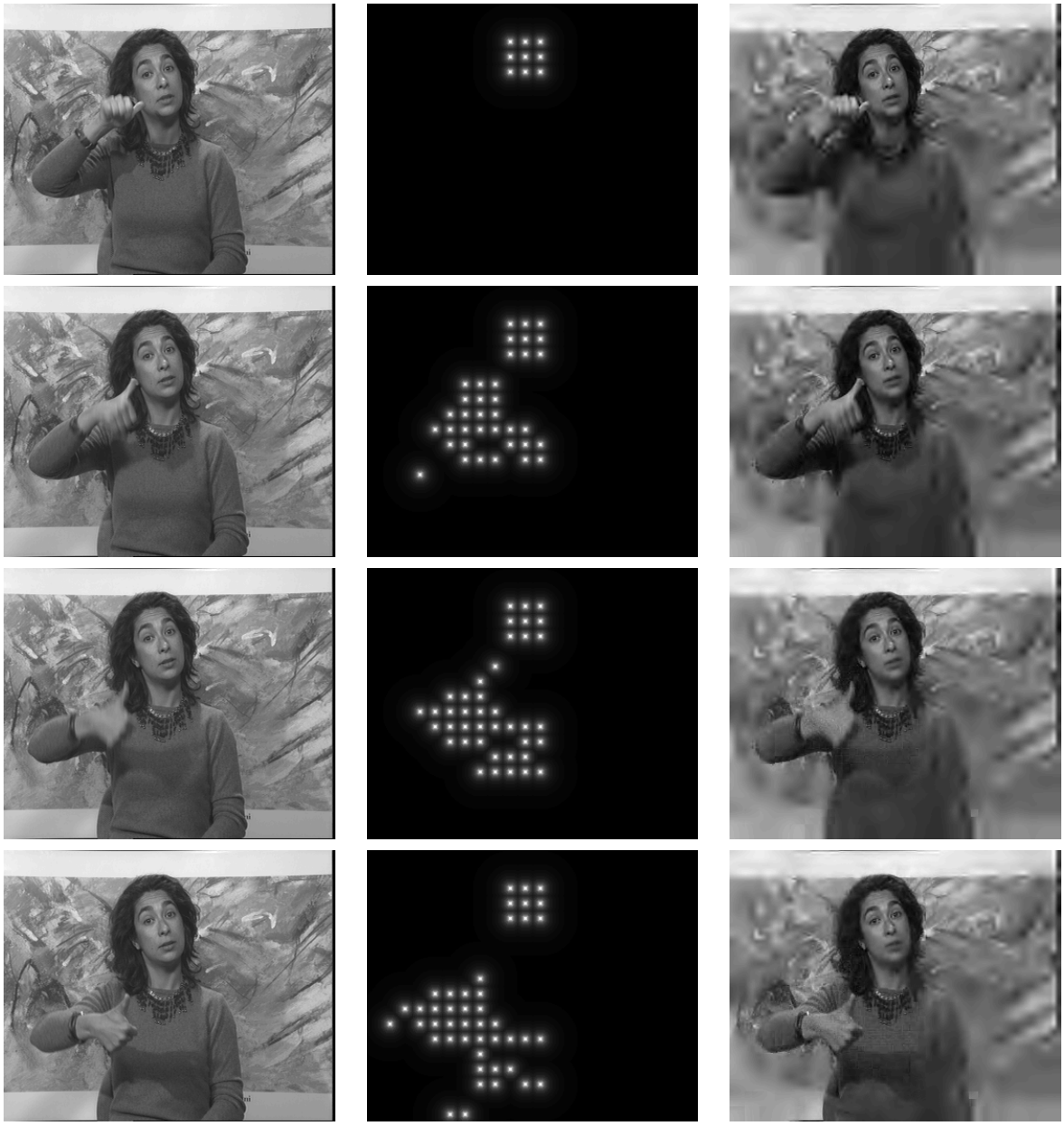

Fig. 10. Consecutive frames of the "Silence" sequence (left); the selected foveation points (middle); and the FSVC compression results at 200 Kbits/sec (right).

coding algorithms being compared include (1) Uniform resolution scalable coding without any foveation models applied; (2) Foveated scalable coding with frame prediction from original previous frames; (3) Foveated scalable coding with frame prediction from base rate coded previous frames; and (4) FSVC with adaptive frame prediction. At the same bit rate of 200 Kbits/sec, FSVC with adaptive frame prediction exhibits the best foveated subjective quality.

To demonstrate the scalable features of FSVC, Fig. 12 compares the FWQI results of the decompressed "Salesman" sequences at 200, 400 and 800 Kbits/sec, respectively. The reconstructed sequences are created from the same encoded bitstream by truncating the bitstream at different places. Fig. 13 shows the reconstructed 32th frame of the "Salesman" sequence at 200, 400 and 800 Kbits/sec, respectively. The results exhibit not only the rate scalable feature but also the *foveation scalable* characteristic of FSVC, in the sense that the foveation depth increases with the decrease of bit rate.

V. Conclusions and Discussions

In this paper, a new wavelet-based scalable foveated video coding system, FSVC, is proposed. A foveation-based HVS model plays an important role in the system. It helps the coding system to foveate on the visually important components in the video sequence. FSVC is a flexible prototype that can incorporate various kinds of foveation

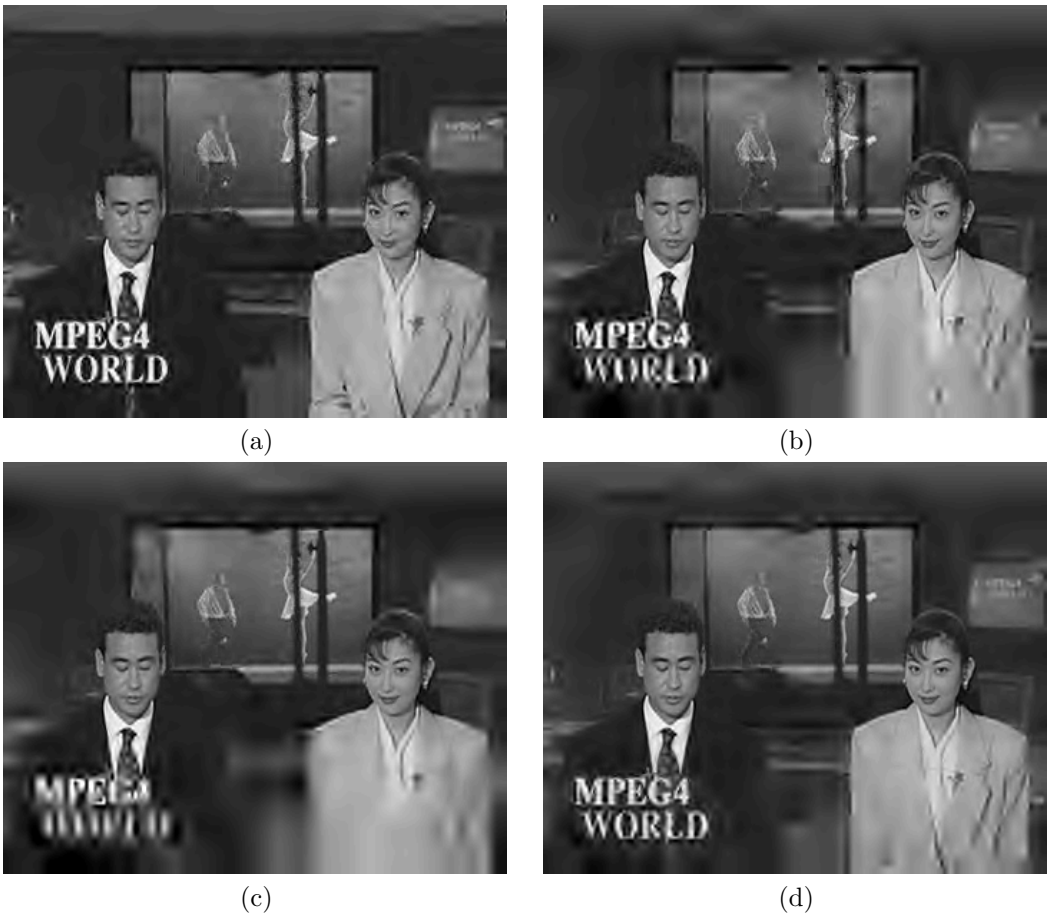

Fig. 11. Frame 26 of the "News" sequence compressed at 200 Kbits/sec with (a) uniform resolution scalable coding; (b) foveated scalable coding with frame prediction from original previous frames; (c) foveated scalable coding with frame prediction from base rate coded previous frames; and (d) FSVC with adaptive frame prediction, respectively.

point selection schemes to fit in different application environments. Specifically, we implemented a foveation region selection algorithm for the encoding of video sequences with human faces. A novel automated foveation point selection scheme and an adaptive frame prediction algorithm is proposed. By using the adaptive frame prediction algorithm, error propagation is well controlled, while at the same time, better frame prediction is achieved.

The FSVC technique has many potential applications. One application is *knowledge-based* video coding. Many different kinds of knowledge about the contents and the contexts of the encoded sequences can be naturally embedded into the general FSVC system. This implies that FSVC is very good for special-purpose video communication applications such as videoconferencing and telemedicine, where a lot of prior information is available to the encoder. If an Audio-Visual Object (AVO) description [2], [4] of the scene is available, then higher visual quality-compression performance can be expected. In general, the more we know about the video signal being encoded, the more we can improve the performance of FSVC.

FSVC is very suitable for *dynamic* variable bit rate network video transmission. For example, if the available bandwidth drops dramatically on the network, a fixed data rate coding system has to stop transmission. A uniform resolution scalable coding system can still work properly but might transmit completely unacceptable quality video to the client. A FSVC system, however, may still deliver useful information to the client, who might be specifically

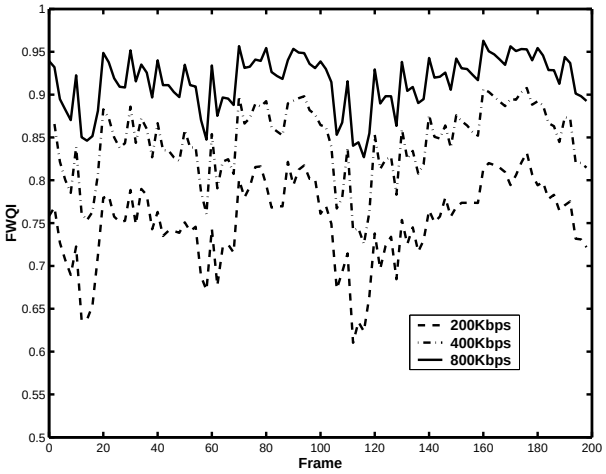

Fig. 12. FWQI measurement results of "Salesman" sequence at 200, 400 and 800 Kbits/sec, respectively.

interested in certain areas in the video frame during each time piece.

FSVC also provides greater flexibility for *multi-user* and *heterogeneous* network video communications. If the video server needs to send video signals to different users with very different bandwidth connections, then FSVC, with only one-time encoding, supports the possibility to provide every user with the best quality video he/she can get in terms of foveated quality measurement.

Finally, FSVC is also a good choice for *interactive* video communications, where the users are involved in giving feedback information to the other side of the communication system. The feedback information may be regions or objects of interest and can be converted into knowledge about the video sequence inside the FSVC encoder. Consequently, improved video quality can be achieved.

References

[1] B. Barnett, "Basic concepts and techniques of video coding and the H.261 standard," in *Handbook of Image and Video Processing* (A. Bovik, ed.), Academic Press, May 2000.

[2] A. Puri and T. Chen, *Multimedia Systems, Standards, and Networks*. New York: Marcel Dekker, Inc., 2000.

[3] S. Arawvith 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.

[4] B. Erol, A. Dumitras, and F. Kossentini, "Emerging MPEG standards: MPEG-4 and MPEG-7," in *Handbook of Image and Video Processing* (A. Bovik, ed.), Academic Press, May 2000.

[5] T. N. Pappas and R. J. Safranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Proc.* (A. Bovik, ed.), Academic Press, 2000.

[6] C. J. van den Branden Lambrecht, Ed., "Special issue on image and video quality metrics," *Signal Proc.*, vol. 70, Nov. 1998.

[7] J. Lubin, "A visual discrimination mode for image system design and evaluation," in *Visual Models for Target Detection and Recognition* (E. Peli, ed.), pp. 207–220, Singapore: World Scientific Publishers, 1995.

[8] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in *Proc. SPIE*, vol. 1616, pp. 2–15, 1992.

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

[10] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Proc.*, vol. 6, pp. 1164–1175, Aug. 1997.

[11] Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Applications* (B. Furht and O. Marques, eds.), CRC Press, 2003.

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

[13] 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 Tech.*, vol. 6, pp. 243–250, June 1996.

[14] D. Taubman, C. Chrysafis, and A. Drukarev, "Embedded block coding with optimized truncation," *ISO/IEC JTC1/SC29/WG1, JPEG-2000 Document WG1N1129*, Nov. 1998.

[15] D. Taubman and A. Zakhor, "Multirate 3-D subband coding of video," *IEEE Trans. Image Proc.*, vol. 3, pp. 572–588, Sept. 1994.

[16] Y. W. Chen and W. A. Pearlman, "Three-dimensional subband coding of video using the zerotree method," in *Proc. SPIE Visual Comm. and Image Processing*, vol. 2727, Mar. 1996.

[17] K. S. Shen and E. J. Delp, "Wavelet based rate scalable video compression," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 9, pp. 109–122, Feb. 1999.

[18] J. Y. Tham, S. Ranganath, and A. A. Kassim, "Highly scalable wavelet-based video codec for very low bit-rate environment," *IEEE Journal on Selected Areas in Comm.*, vol. 16, pp. 12–27, Jan. 1998.

[19] S.-J. Choi and J. W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. Image Proc.*, vol. 8, pp. 155–167, Feb. 1999.

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

[21] J.-Y. Lee, H.-S. Oh, and S.-J. Ko, "Motion-compensated layered video coding for playback scalability," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 11, pp. 619–628, May 2001.

[22] H. Sun, W. Kwok, and J. W. Zdepski, "Architectures for MPEG compressed bitstream scaling," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 6, pp. 191–199, Apr. 1996.

[23] P. A. A. Assuncao and M. Ghanbari, "A frequency-domain video transcoder for dynamic bit-rate reduction of MPEG-2 bit streams," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 8, pp. 953–967, Dec. 1998.

[24] 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.

[25] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA Engineer*, vol. 29, no. 6, pp. 33–41, 1984.

[26] C. Bandera and P. Scott, "Foveal machine vision systems," in *Proc. IEEE Int. Conf. System, Man and Cybernetics*, pp. 596–599, Nov. 1989.

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

[28] 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.

[29] P. Kortum and W. S. Geisler, "Implementation of a foveal image coding system for image bandwidth reduction," in *Proc. SPIE*, vol. 2657, pp. 350–360, 1996.

[30] P. Camacho, F. Arrebola, and F. Sandoval, "Shifted fovea multiresolution geometries," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 1, pp. 307–310, 1996.

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

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

[33] 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, pp. 235–243, Mar. 1999.

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

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

[36] H. R. Sheikh, S. Liu, B. L. Evans, and A. C. Bovik, "Real-time foveation techniques for H.263 video encoding in software," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, vol. 3, pp. 1781–1784, May 2001.

[37] E. Nguyen, C. Labit, and J.-M. Odobez, "A ROI approach for hybrid image sequence coding," in *Proc. IEEE Int. Conf. Image Proc.*, pp. 245–249, 1994.

[38] A. E. Yagle, "Region-of-interest tomography using the wavelet transform and angular harmonics," *IEEE Signal Processing Letters*, vol. 1, pp. 134–135, Sept. 1994.

[39] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, "Low bit-rate coding of image sequences using adaptive regions of interest," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 8, pp. 928–934, Dec. 1998.

[40] D. Nister and C. Christopoulos, "Lossless region of interest coding," *Signal Proc.*, vol. 78, pp. 1–17, Oct. 1999.

[41] Z. Wang, *Rate scalable foveated image and video communications*. PhD thesis, Dept. of ECE, The University of Texas at Austin, Dec. 2001.

[42] E.-C. Chang, *Foveation techniques and scheduling issues in thin-wire visualization*. PhD thesis, Dept. of CS, New York University, 1998.

Fig. 13. 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.

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

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

[45] Z. Xiong and K. Ramchandran, "Wavelet image compression," in *Handbook of Image and Video Processing* (A. Bovik, ed.), Academic Press, May 2000.

[46] C. Podilchuk, N. Jayant, and N. Farvardin, "Three-dimensional subband coding of video," *IEEE Trans. Image Proc.*, vol. 4, pp. 125–139, Feb. 1995.

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

[48] S. G. Mallat, "Multifrequency channel decomposition of images and wavelet models," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol. 37, pp. 2091–2110, 1989.

[49] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using the wavelet transform," *IEEE Trans. Image Proc.*, vol. 1, pp. 205–220, Apr. 1992.

[50] R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, pp. 1945–1948, May 1989.

[51] I. Hontsch, L. J. Karam, and R. J. Safranek, "A perceptually tuned embedded zerotree image coder," in *Proc. IEEE Int. Conf. Image Proc.*, pp. 41–44, Oct. 1997.

[52] T. O. Salmon, "Fixational eye movement." *VS III: Ocular Motility and Binocular Vision*, College of Optometry, Northeastern State University, 2001.

[53] K. Sayood, *Introduction to Data Compression*. San Francisco: Morgan Kaufmann Publishers, Inc., 1996.

[54] 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.

[55] 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.

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

[57] 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 Tech.*, vol. 7, pp. 615–628, Aug. 1997.

[58] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Trans. Multimedia*, vol. 1, pp. 264–277, Sept. 1999.

[59] K. S. Shen and E. J. Delp, "A control scheme for a data rate scalable video codec," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 2, pp. 69–72, Sept. 1996.

[60] Z. Wang, A. C. Bovik, and L. Lu, "Wavelet-based foveated image quality measurement for region of interest image coding," in *Proc. IEEE Int. Conf. Image Proc.*, Oct. 2001.

[61] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, pp. 81–84, Mar. 2002.

[62] Z. Wang, "Demo images and free software for 'a universal image quality index'," `http://anchovy.ece.utexas.edu/~zwang/research/quality_index/demo.html`.

[63] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, vol. 4, (Orlando), pp. 3313–3316, May 2002.