RATE SCALABLE VIDEO CODING USING A FOVEATION-BASED HUMAN VISUAL SYSTEM MODEL

Zhou Wang[1], *Ligang Lu*[2], *and Alan C. Bovik*[1]

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

ABSTRACT

Recently, there are two interesting trends in image and video coding research. One is to use human visual system (HVS) models to improve the current state-of-the-art coding algorithms by better exploiting the properties of the intended receiver. The other is to design rate scalable video codecs, which allow the extraction of coded visual information at continuously varying bit rates from a single compressed bitstream. In this paper, we follow these two trends and propose a foveation scalable video coding (FSVC) algorithm, which supplies good quality-compression performance as well as effective rate scalability to support simple and precise bit rate control. A foveation-based HVS model plays a key role in the algorithm. The algorithm is amenable to the inclusion of various HVS models and adaptable to different video communication applications.

1. INTRODUCTION

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. Currently, peak-signal-to-noise-ratio (PSNR) is still widely employed to test image and video quality. However, it is well accepted that perceived image and video quality does not correlate well with PSNR. Human visual system (HVS) characteristics must be considered [1, 2]. Although the current understanding of the HVS is still insufficient to provide a precise, generic and robust algorithm to measure perceived image and 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 coding techniques.

Rate scalable coding algorithms allow the extraction of coded visual information at continuously varying data rates from a single compressed bitstream. This feature is especially suited for video transmission over heterogeneous, multi-user, time-varying and interactive networks such as the Internet. For example, in order to provide video services over the Internet, the video server must have the ability to create variable bandwidth video streams to meet different user requirements. The traditional solutions, such as layered video, video transcoding, 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 the 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. In contrast, with a rate scalable codec, we can tightly couple the available bandwidth and the data rate of the video being delivered. Recently, a class of embedded coding algorithms has received great attention. The well-known algorithms include the embedded zero tree wavelet (EZW) algorithm [3], and the set partitioning in hierarchical trees (SPIHT) algorithm [4]. Embedded wavelet coding not only provides good coding performance, but also has the property of rate scalability. Many recent image and video coding algorithms are developed based on the idea originated from EZW.

In this paper, we propose a new wavelet-based video coding technique called foveation scalable video coding (FSVC), which is a highly rate scalable video coding method that attempts to optimize visual quality at arbitrary bit rate within the bandwidth range. A foveation-based HVS model is at the core of the algorithm, which is used for multiple purposes including adaptive importance weighting of the wavelet coefficients and adaptive frame prediction for motion compensation.

2. GENERAL FRAMEWORK

Similar to many other video coding algorithms, 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). Fig. 1 gives the general framework for the encoding of I and P frames.

The encoding of the I frame is equivalent to the encoding of a still image. The method is based on our algorithm introduced in [5]. We first apply the discrete wavelet transform (DWT) to the image and obtain the wavelet coefficients. A foveation-baesd HVS model is employed to determine the visual importance of the wavelet coefficients and to give each wavelet coefficient an important weighting value, which is then used to weight the wavelet coefficient. An embedded encoding algorithm is employed to generate the rate scalable bitstream.

In the encoding of the P frames, motion estimation (ME) and motion compensation (MC) techniques are employed to exploit temporal redundancy. Frame prediction plays an important role in ME/MC based algorithm. Different from previous algorithms, FSVC uses two instead of one version of the previous frames to do the prediction. One is the original previous frame. The other is a feedback base-rate decoded version of the previous frame. The combination is based on the foveation weighting model.

The HVS modeling methods are different for I frames and P frames. During the encoding process, a rate control algorithm is

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

used to allocate bits to each frame. The allocation is determined by the available bandwidth, the user requirements, the HVS modeling and the frame prediction error.

3. FOVEATION-BASED HVS MODEL

3.1 Foveated Visual Sensitivity Model

Let us first examine the anatomy of the early vision system. The light first passes through the optics of the eye and is then sampled by the photoreceptors (cones and rods) on the retina. The cone receptor distribution is highly non-uniform. The photoreceptors deliver data to the bipolar cells, which in turn supply information to the ganglion cells, which also have a highly non-uniform distribution. The variation of the densities of photoreceptors and ganglion cells with eccentricity is shown in Fig. 2.

The densities of cones and ganglion cells play important roles in determining the ability of our eyes to resolve what we see. Psychovisual experiments had be conducted to measure the contrast sensitivity as a function of retinal eccentricity. In [6], the model that fits the experimental data is given by

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

where f is the spatial frequency (cycles/degree), e is the retinal eccentricity (degrees), CT_0 is a constant minimal contrast threshold, α is the spatial frequency decay constant, e_2 is the half-resolution eccentricity, and $CT(f, e)$ is the visible contrast threshold as a function of f and e. The best fitting parameter values given in [6] are $\alpha = 0.106$, $e_2 = 2.3$, and $CT_0 = 1/64$. The contrast sensitivity is defined as:

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

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

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

Given a pixel $\boldsymbol{x}$ in an N pixels wide image, its distance from the foveation point $\boldsymbol{x}_f$ is $d(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}^f\|_2$ (pixels) and its eccentricity is given by $e(v, \boldsymbol{x}) = \tan^{-1}\left(d(\boldsymbol{x})/Nv\right)$, where v is the viewing distance in image width. The maximum perceived resolution is also limited by the display resolution $r \approx \pi Nv/180$ (pixels/ degree). The Nyquist display frequency is given by $f_d = r/2$ (cycles/degree). Combining this with (3), the cutoff frequency for $\boldsymbol{x}$ is $f_m(\boldsymbol{x}) = \min(f_c(d(\boldsymbol{x})), f_d)$. We define the foveation-based error sensitivity as:

$$S_f(v, f, \boldsymbol{x}) = \begin{cases} \dfrac{CS(f, e(v,\boldsymbol{x}))}{CS(f,0)} & for \quad f \le f_m(\boldsymbol{x}) \\ 0 & for \quad f > f_m(\boldsymbol{x}) \end{cases} \tag{4}$$

Many other HVS features are also related to perceived video quality. We are especially interested in the foveation feature not only because it is a very promising feature to effectively remove information redundancy from peripheral regions, but also because it makes it possible to establish a generalized foveation-based framework to embrace the other HVS features.

3.2 Foveation Point(s) Selection

Psychological experiments show that statistically, the human eyes' fixation points are very non-uniformly distributed [7, 8]. Depending on the applications, foveation point(s) selection can be done by an automatic or an interactive method. We are interested in the automatic method in this paper. Although only one foveation point exists at one time for one human observer, it is necessary to allow multiple foveation points, because there may exist multiple points that have a high possibility of attracting attention and there may be multiple observers at the same time. It is also reasonable to allocate foveation points at the areas where the HVS is very sensitive to errors. This is actually a generalization of the foveation concept.

In FSVC, we partition the whole picture into blocks. The candidate foveation points are the centers of all these blocks. For I frames, we first determine the regions of interest (ROIs) and put foveation points inside those regions. Specifically, we regard the face regions as the ROIs and use a face detection algorithm similar to that in [9] to find faces. The light adaptation feature of the HVS is also considered. For P frames, the foveation points are chosen in the regions where the prediction errors are larger than a threshold. We care more about the ROIs and use a smaller prediction error threshold there. Examples are given in Fig. 5.

3.3 HVS-Based Weighting Model in the DWT Domain

In embedded wavelet image coding, a multilevel 2-D DWT is applied to decompose the original image into subbands. A typical DWT decomposition structure is shown in Fig. 3(a). We use (λ, θ) to denote the subband at decomposition level λ and orientation θ, where θ can be LL, LH, HH, or HL. Our algorithm assigns each wavelet coefficient an importance weight. The embedded coding algorithm is then applied to the weighted wavelet coefficients, so that the visually important coefficients and bits are encoded and transmitted earlier.

The wavelet coefficients at different subbands supply information regarding variable perceptual importance. In [10], the visual sensitivity is measured in wavelet decompositions. We define the error sensitivity at subband (λ, θ) as:

$$S_w(\lambda,\theta) = 1/T_{\lambda,\theta} \tag{5}$$

where $T_{\lambda,\theta}$ is the error detection threshold at subband (λ,θ) as given in [10]. Within each wavelet subband, we first find its corresponding foveation point. For a given coefficient at position $\boldsymbol{x}$, residing in subband (λ, θ), it is easy to calculate its equivalent distance $\boldsymbol{d}_{\lambda,\theta}(\boldsymbol{x})$ from the foveation point in the spatial domain. Given the spatial frequency of the subband $f = r2^{-\lambda}$ [10],

$$S_f(v, f, \boldsymbol{x}) = S_f(v, r2^{-\lambda}, d_{\lambda,\theta}(\boldsymbol{x})). \tag{6}$$

A foveation-based HVS error sensitivity model in the DWT domain is obtained by combining (5) and (6):

$$S(v,\boldsymbol{x}) = \left[S_w(\lambda,\theta)\right]^{\beta_1} \cdot \left[S_f(v, r2^{-\lambda}, d_{\lambda,\theta}(\boldsymbol{x}))\right]^{\beta_2}, \tag{7}$$

where β_1 and β_2 are parameters used to control the magnitudes of S_w and S_f, respectively. If we have P (P>1) foveation points, we calculate $S_i(v,\boldsymbol{x})$ for i = 1, 2, …, P. The error sensitivity is then given by $S(v,\boldsymbol{x}) = \max[S_i(v,\boldsymbol{x})]$.

In practice, v is unknown to us. In order to assign each coefficient a fixed weight, we assume a probability distribution $p(v)$ [5] of the viewing distance and the weighting is given by

$$W(\boldsymbol{x}) = \int_{0^+}^{\infty} p(v) S(v,\boldsymbol{x}) dv \tag{8}$$

Fig. 3(b) gives an example of the weighting model, where a circular foveation region is located around the face region. Applying this model to the "Zelda" image, we obtain the results given in Fig. 4. Examples of the I frame and P frame foveation-based HVS weighting are shown in Fig. 5.

4. IMPLEMENTATION

The dynamic range of the wavelet coefficients is largely expanded after weighting. This makes the original EZW and SPIHT encoding less efficient. Therefore, we use a modified SPIHT algorithm [5] to do the embedded encoding.

Frame prediction plays an important role in ME/MC based video coding. This task is more challenging in rate scalable video coding than in fixed-rate coding, because the decoding bit rate is unknown to the encoder. A simple method is to use the original motion compensated frame as the prediction, but this leads to poor prediction and the errors will propagate to the following frames. Another method [11] is to use a low base bit rate decoded and motion compensated frame as the prediction. This method avoids the significant error propagation problem. However, large prediction errors occur when the decoding bit rate is much higher than the base bit rate. We use a new method in FSVC, where the original motion compensated frame and the base bit rate motion compensated frame are combined to make the prediction. The combination is based on the foveation-based weighting model. For the regions around the foveation points, more weight is given to the base bit rate motion compensated reference frames, while for the regions far from the foveation points, more weight is given to the high quality motion compensated reference frames. By using the new algorithm, error propagation becomes small, while at the same time, better frame prediction is achieved, which leads to smaller prediction error and better compression performance. More details about the new frame prediction method are given in [12].

We tested our algorithm on CIF size, YCbCr 4:2:0 format video sequences. For general video sequences, comparable visual quality to MPEG and H.263 codecs is obtained at 500Kbps or higher rates. For face video sequences such as 'News', enhanced visual quality is achieved.

5. CONCLUSION

We propose a new video coding scheme called FSVC, which combines a foveation-based HVS model with the embedded wavelet coding technique, so that the output bits are ordered according to visual importance. FSVC can easily adapt to various HVS models. It has many potential applications, such as knowledge-based video coding, and video transmission over heterogeneous, multi-user, time-varying, and interactive networks.

REFERENCES

[1] T. N. Pappas, and R. J. Saranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Processing*, Al Bovik, ed., Academic Press, 2000.

[2] B. Girod, "What's wrong with Mean-Squared Error," *Digital Images and Human Vision*, Chapter 15, A. B. Watson, Ed., 1993.

[3] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, no. 3445-3462, Dec. 1993.

[4] 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, no. 3, pp. 243-250, June 1996.

[5] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Trans. Image Processing*, submitted, 2000.

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

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

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

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

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

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

[12] L. Lu, Z. Wang, and A. C. Bovik, "Adaptive frame prediction for foveation scalable video coding," submitted to *IEEE Inter. Conference on Multimedia and Expo*, 2001.

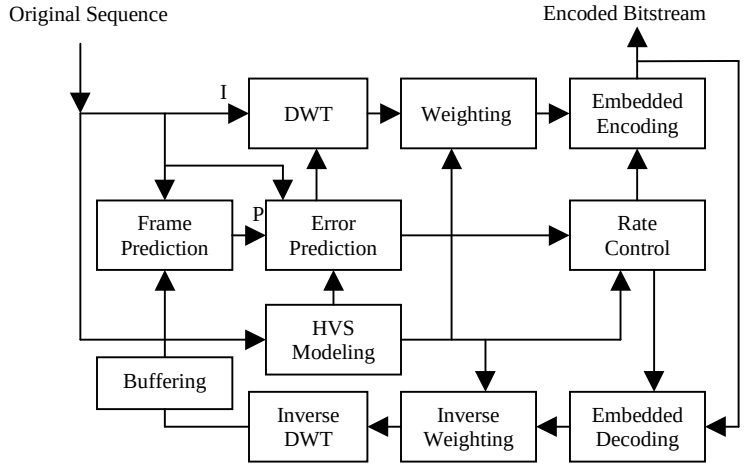


Fig. 1 General framework of the FSVC encoding system.

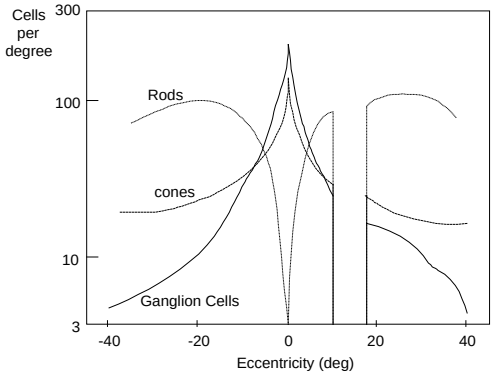
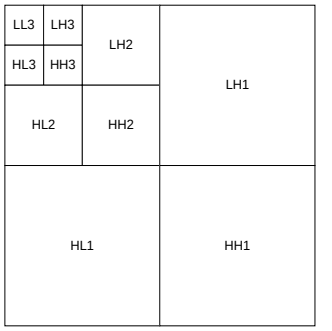
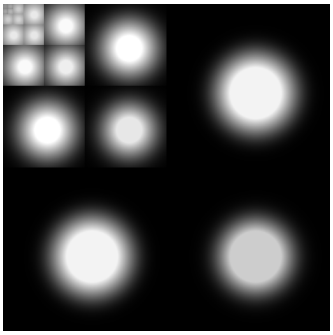


Fig. 2 Photoreceptor and ganglion cell distribution in human retina.



(a)



(b)

Fig. 3 (a) The DWT decomposition structure; (b) Foveation-based HVS weighting model in the DWT domain.

Fig. 4 Foveated rate scalable compression results of "Zelda" image (Upper-left, 512×512, 8bits/pixel) using the weighting model in Fig. 3(b). The compression ratios are 1024:1 (Upper-right), 256:1 (Lower-left) and 32:1 (Lower-right), respectively.

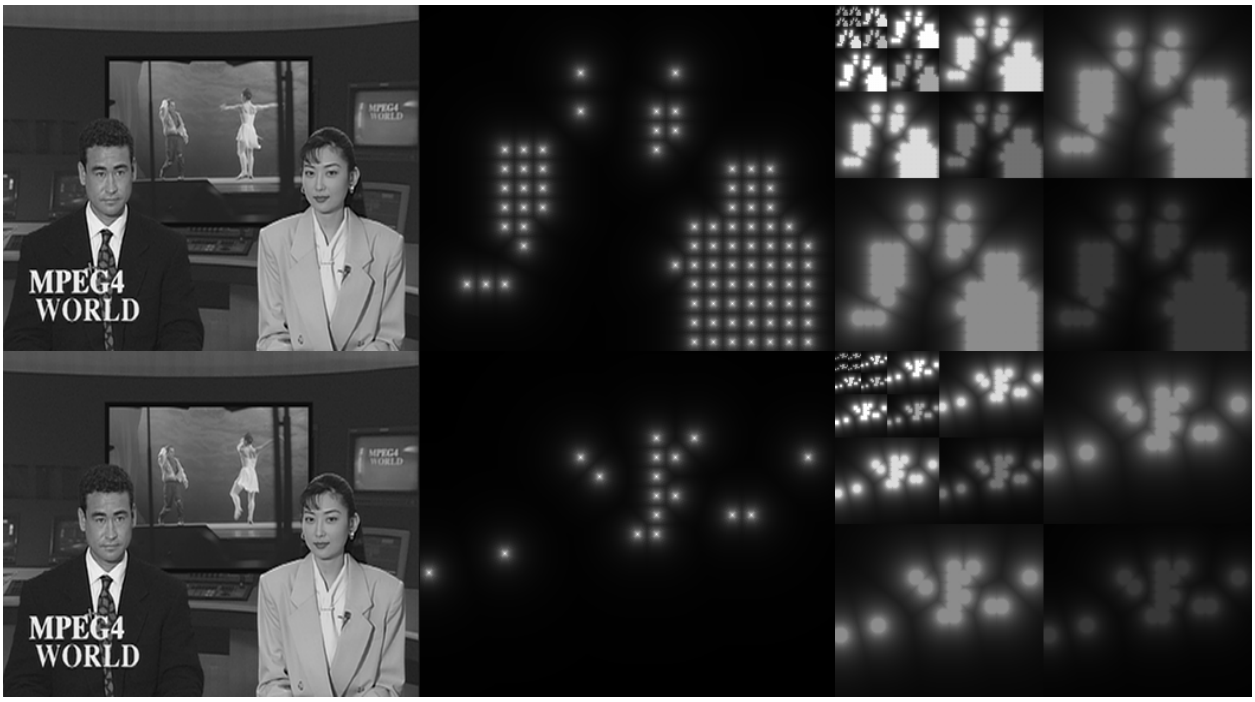


Fig. 5 I frame and P frame foveation point selection results and DWT domain weighting models. Upper-left: an I frame; Bottom-left: a P frame; Upper-center: Selected foveation points for the I frame; Bottom-right: Selected foveation points for the P frame; Upper-right: DWT domain weighting model for the I frame; Bottom-right: DWT domain weighting model for the P frame.