ADAPTIVE FRAME PREDICTION FOR FOVEATION SCALABLE VIDEO CODING

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

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

ABSTRACT

Embedded rate scalable video coding allows for the extraction of coded visual information at continuously varying bit rates from a single compressed bitstream. This is a very attractive feature for many multimedia communication applications. Motion estimation (ME) /motion compensation (MC) techniques are widely employed in various video coding systems to reduce temporal information redundancy. One of the major challenging problems in ME/MC based rate scalable video coding is how to generate the prediction frame from the previous frame to match the current frame. This problem is more difficult in rate scalable coding than in fixed rate coding because the decoding data rate is unavailable to the encoder. We propose an adaptive frame prediction scheme for foveation scalable video coding (FSVC), which is a new video coding algorithm that combines a foveation-based human visual system (HVS) model with a wavelet-based rate scalable coding algorithm. The new frame prediction algorithm provides an adaptive mechanism to control the prediction errors while reduce error propagation.

1. INTRODUCTION

Recently, there have emerged two interesting research trends that are very promising and may lead to significantly better video codecs in the near future. The first trend is to embed human visual system (HVS) models into the coding system. It is well accepted that perceived image and video quality does not correlate well with peak signal-to-noise-ratio (PSNR) [1], which is still widely used to test video quality. HVS characteristics must be considered to provide better visual quality measurements. Although the current understanding of the HVS still is 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 image and video coding algorithms. Various HVS features can be employed to improve the video coding systems. The most widely used characteristics include the spatial texture masking effect, and the perceptual variances of the spatial and temporal frequency sensitivities. Another useful human visual characteristic is the foveation feature, which stands for the fact that the visual resolution of the HVS is highest around the point of fixiation (foveation point) and decreases rapidly with the increasing eccentricity. The second recent trend in video coding research is to develop rate scalable coding techniques [2-4], which allow for 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, dynamic and interactive networks. Following these two trends, we proposed [5] a new wavelet-based video coding scheme called foveation scalable vidoe coding (FSVC), which is a highly rate scalable video coding method that attempts to optimize foveated visual quality at arbitrary bit rate within the bandwidth range. A foveation-based HVS model plays an important role in the system. More information about how the model is developed and how it is used for image and video coding is given in [5-7].

Motion estimation (ME) / motion compensation (MC) techniques are widely employed in many video coding systems to reduce temporal information redundancy. In ME/MC based video coding algorithms, a prediction frame is generated from the previous reference frame via ME/MC to estimate the current frame. The prediction error frame, instead of the original current frame, is encoded. If good prediction is made, then the prediction error is small and thus better compression is expected. At the decoder side, the decoded current frame is achieved by adding the prediction frame obtained from MC and the decoded prediction error frame. One of the major challenges in ME/MC based continuously rate scalable video coding is how to do frame prediction. It is more complicated than fixed rate coding because the decoding data rate is unknown to the encoder. This leads to inconsistent reference frames at the encoder and the decoder. The inconsistency is a source of error and the error may propagate to the frames that follow. In [2, 3], a control mechanism is proposed to avoid the error propagation problem with the sacrifice of prediction precision and coding efficiency.

The main purpose of this paper is to propose an adaptive frame prediction algorithm for FSVC. In Section 2, we give a brief review of FSVC. Section 3 discusses the methods to solve the frame prediction problem and introduces our new adaptive prediction method. Finally, Section 4 concludes the paper.

2. REVIEW OF FOVEATION SCALABLE VIDEO CODING SYSTEM

Based on some recently published psychovisual research results on HVS's foveation feature, we proposed a foveation-based visual sensitivity model [6, 7]. An example is shown in Figure 1. With this model, a foveation-based HVS weighting mask can be calculated in both spatial and discrete wavelet transform (DWT) domain. The DWT decomposition architecture and an example of

This work was partially supported by IBM Corporation, Texas Instrument, Inc., and Texas Advanced Technology Program.

the DWT domain weighting mask is shown in Figure 2 [7]. In practice, it is necessary to allow multiple foveation points because there may exist multiple points attracting the human observer's attention and there may be multiple users watching the video simultaneously. It is also reasonable to set the foveation points at areas where the human eyes are very sensitive to errors. This is actually an extension of the foveation concept.

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). The general framework for the encoding of I frames and P frames is shown in Figure 3. For an I frame, we first apply DWT and obtain the wavelet coefficients. The foveation-based HVS weighting mask is computed and employed to weight the visual importance of the wavelet coefficients. Embedded bit plane coding algorithms such as SPIHT [8] have been widely adopted for rate scalable image and video coding. We use a modified SPIHT algorithm [7] to encode the weighted wavelet coefficients. For the encoding of the P frames, two instead of one version of the previous frames are employed to generate the reference frame. One is the original previous frame. The other is a feedback decoded version of the previous frame. This is different from other video coding algorithms. Block-based motion estimation is applied and the reference frame is motion compensated on a block basis and subtracted from the original current frame to create the prediction error frame. The DWT is then applied to the prediction error frame, and the resulting coefficients are weighted and encoded with the modified SPIHT.

The methods to select foveation points for I frames and P frames are different. For I frames, we put the foveation points at the regions of interest (ROIs). Specifically, we set the face regions as the ROIs and a face detection algorithm is employed. A different strategy is used for the foveation point selection of P frames. The reason is that P frames are not encoded directly, but rely on their previous frames. Only the difference between the current frame and the prediction from the previous frame is of interest to us. FSVC mainly selects foveation points in the regions whose prediction errors are larger than a threshold value. Since human observers' attention is very likely to fixate on ROIs, even very small movements or changes there are likely to be noticed. Therefore, we use a smaller threshold in the ROIs, thus small changes in ROIs will generate foveation points. In Figure 4, an I frame in the "News" sequence, the following P frame, and the spatial domain weighting mask created from the foveation point selection of the P frame are shown.

3. 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 prediction of the current frame. With this choice, the prediction frames are exactly the same at the encoder and the decoder. However, this choice is impossible 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 of very large difference. The consequence is that very good frame prediction at the encoder side may produce bad prediction at the decoder side. In addition, the bad prediction error will propagate to all the follwing 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 had been used in [2, 3]. 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 short, this solution results in less precise prediction and less efficient compression.

Our adaptive prediction algorithm is 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 with the foveation-based HVS model. The new frame prediction algorithms are shown in Figure 5. 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 $W_S(\boldsymbol{x})$ be the normalized weight at location $\boldsymbol{x}$, where the value range of $W_S(\boldsymbol{x})$ is [0, 1]. Let $I_O(\boldsymbol{x})$ and $I_B(\boldsymbol{x})$ be the pixel values at location $\boldsymbol{x}$ of the motion compensated original reference frame and base rate decoded reference frame, respectively. Then the combined encoder prediction value $I_E(\boldsymbol{x})$ is given by:

$$I_E(\boldsymbol{x}) = [1 - W_S(\boldsymbol{x})]I_O(\boldsymbol{x}) + W_S(\boldsymbol{x})I_B(\boldsymbol{x}). \qquad (1)$$

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 $I_C(\boldsymbol{x})$ be the pixel values at location $\boldsymbol{x}$ of the motion compensated reference frame at the current decoding bit rate, then the combined decoder prediction value $I_D(\boldsymbol{x})$ is:

$$I_D(\boldsymbol{x}) = [1 - W_S(\boldsymbol{x})]I_C(\boldsymbol{x}) + W_S(\boldsymbol{x})I_B(\boldsymbol{x}). \qquad (2)$$

The idea behind the weighting equations (1) and (2) 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. The frame predictions at the encoder and decoder are not exactly the same. Subtracting (2) from (1) yields $I_E(\boldsymbol{x}) - I_D(\boldsymbol{x}) = [1 - W_S(\boldsymbol{x})][I_O(\boldsymbol{x}) - I_C(\boldsymbol{x})]$. Since at the difficult prediction regions, the value of $W_S(\boldsymbol{x})$ is large (usually very close or equal to 1), the error between $I_E(\boldsymbol{x})$ and $I_D(\boldsymbol{x})$ is very small and can be neglected. At the easy

prediction regions, the values of $I_C(\boldsymbol{x})$ is very close to $I_O(\boldsymbol{x})$. Therefore, the prediction difference at the encoder and the decoder is small. By this way, the error propagation is well controlled. Also note that at the easy prediction regions, the value of $W_S(\boldsymbol{x})$ is small and the actual prediction in (1) and (2) is mainly from $I_O(\boldsymbol{x})$ and $I_C(\boldsymbol{x})$. Since $I_O(\boldsymbol{x})$ and $I_C(\boldsymbol{x})$ are from high quality prediction frames, their prediction values are much better than the poor prediction of $I_B(\boldsymbol{x})$. By this way, the prediction errors are reduced.

4. CONCLUSIONS

Frame prediction is a very challenging problem in ME/MC based rate scalable video coding. We proposed an adaptive frame prediction algorithm for FSVC. By using the new frame prediction algorithm, error propagation is well controlled, while at the same time, better frame prediction is achieved.

5. REFERENCES

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

[2] K. 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.

[3] K. Shen, and E. J. Delp, "A control scheme for a data rate scalable video codec," *IEEE International Conference on Image Processing*, vol.2, pp. 69 –72, 1996.

[4] B.-J. Kim, Z. Xiong, and W. A. Pearlman, "Low bit-rate scalable video coding with 3-D set partitioning in hierarchical trees (3-D SPIHT)," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 10, no.8, pp. 1374-1387, 2000.

[5] Z. Wang, L. Lu, and A. C. Bovik, "Rate scalable video coding using a foveation-based human visual system model," submitted to *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2001.

[6] Z. Wang, A. C. Bovik, and L. Lu, "Wavelet-based foveated image quality measurement for region of interest image coding," submitted to *IEEE International Conference on Image Processing*, 2001.

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

[8] A. Said, and W. A. Pearlman, "A new, fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits & Systems for Video Tech.*, vol. 6, no. 3, pp. 243-250, June 1996.

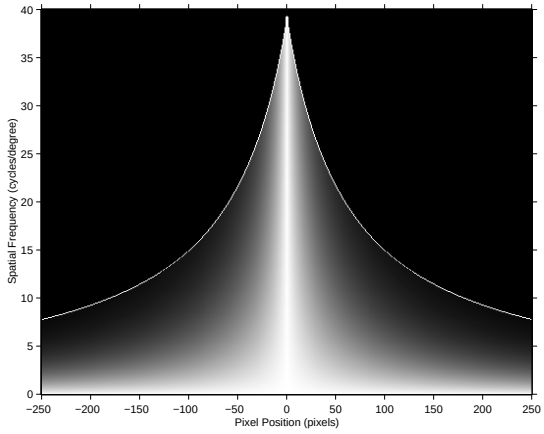

Figure 1. Foveated HVS visual sensitivity model.

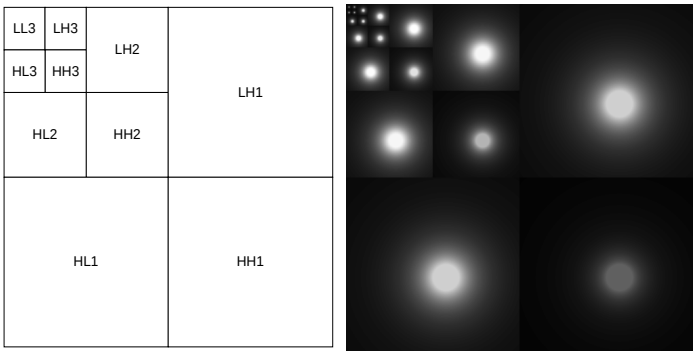

Figure 2. Left: DWT decomposition; Right: Foveation-based HVS weighting mask in the DWT domain.

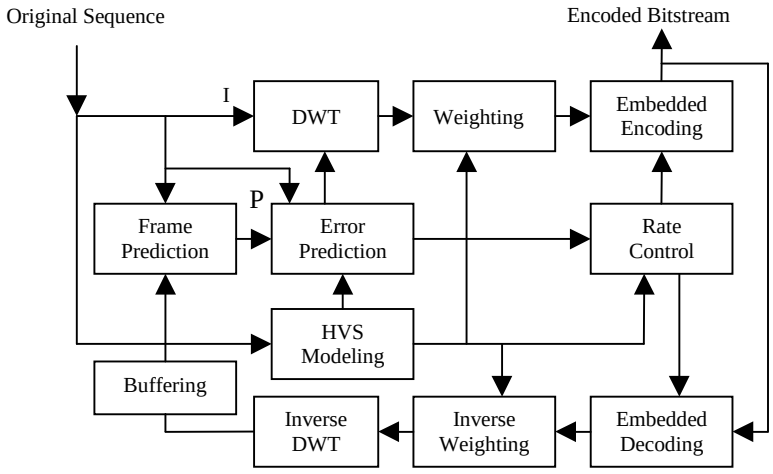

Figure 3. The FSVC encoding system.

Figure 4. Left: an I frame in the "News" sequence; Center: the following P frame; Right: the foveated weighing mask of the P frame.

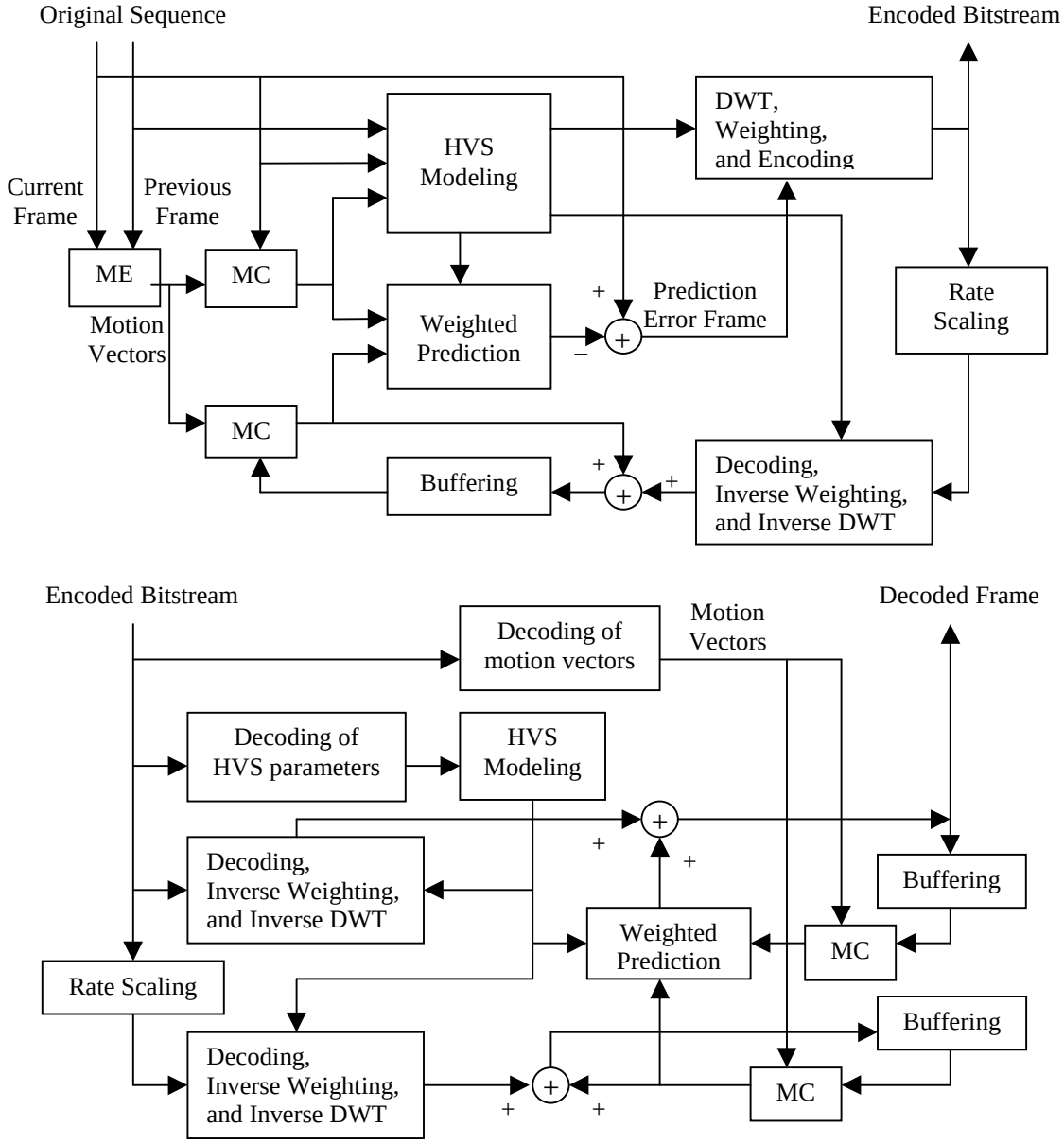

Figure 5. The proposed frame prediction algorithm. Top: The encoder side; Bottom: The decoder side.