Video Coding with Adaptive Motion-Compensated Orthogonal Transforms

Du Liu and Markus Flierl
School of Electrical Engineering
KTH Royal Institute of Technology
Stockholm, Sweden
{dul, mflierl}@kth.se

***Abstract*— Well-known standard hybrid coding techniques utilize the concept of motion-compensated predictive coding in a closed-loop. The resulting coding dependencies are a major challenge for packet-based networks like the Internet. On the other hand, subband coding techniques avoid the dependencies of predictive coding and are able to generate video streams that better match packet-based networks. An interesting class for subband coding is the so-called motion-compensated orthogonal transform. It generates orthogonal subband coefficients for arbitrary underlying motion fields. In this paper, a theoretical signal model based on Gaussian distributions is discussed to construct a cost function for efficient rate allocation. Additionally, a rate-distortion efficient video coding scheme is developed that takes advantage of motion-compensated orthogonal transforms. The scheme combines multiple types of motion-compensated orthogonal transforms, variable block sizes, and half-pel accurate motion compensation. The experimental results show that this adaptive scheme outperforms individual motion-compensated orthogonal transforms by up to 2 dB.**

I. Introduction

The standard video compression techniques, such as H.264/AVC, utilize the concept of motion-compensated predictive coding. Predicted frames (known as P-frames) and bi-predicted frames (B-frames) are used to exploit the temporal redundancy of the sequences with one key frame (I-frame) for each group of pictures (GOP). Because predictive coding is developed in a closed-loop fashion, the coded videos heavily depend on the relationship among successive pictures. These dependencies introduce the risk of error propagation to subsequently decoded pictures, which might be suboptimal in packet loss environments [1]. On the other hand, the motion-compensated orthogonal transform (MCOT) is a subband coding technique that operates in an open-loop fashion [2]. It is an motion adaptive transform that does not depend on predictive coding and, therefore, avoids error propagation. Thus, it is more suitable for packet-based networks like the Internet [3]. The motion-compensated orthogonal transform generates orthogonal subband coefficients for arbitrary underlying motion fields.

Our goal is to develop a rate-distortion efficient video coding scheme that takes advantage of several types of motion-compensated orthogonal transforms. A theoretical transform coding model is discussed to construct a cost function for the type selection. The performance of the practical system will be evaluated by the peak signal-to-noise ratio (PSNR).

This paper is organized as follows: Section II discusses a theoretical signal model for motion-compensated orthogonal transforms. Section III describes the implemented video coding system. The experimental results for the coding system are presented in Section IV. Section V gives a short conclusion.

II. Theoretical Signal Model

Fig. 1 depicts a single MCOT. Assume there are two input pictures $\mathbf{x}_0$ and $\mathbf{x}_1$. Because the two successive pictures are usually similar, $\mathbf{x}_0$ and $\mathbf{x}_1$ can be viewed as a clean picture $\mathbf{v}$ plus independent additive white Gaussian noises n_0 and n_1, respectively. The noise n_0 and n_1 are statistically independent and have the same variance. After the transform, the output signals have one temporal low band L and one energy removed temporal high band H. However, the energy of the noises n_0 and n_1 cannot be shifted after the transform. Thus the temporal subbands are composed by the clean subband signals plus the noises: $L = L_{clean} + \tilde{n}_0$ and $H = H_{clean} + \tilde{n}_1$.

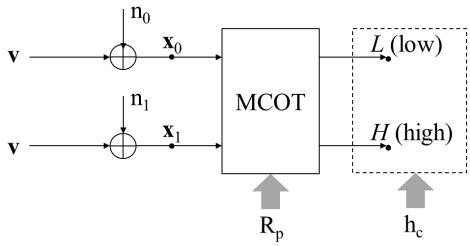

Fig. 1. Signal model for an adaptive orthogonal transform.

Since we would like to describe the performance of the transform, we use the parameter rate R_p to determine how much energy is moved from the high band to the low band, as shown in Fig. 2. R_p indicates the rate of the parameters that are used to perform the adaptive transform. For the practical coding system as introduced in Sec. III, R_p includes the information that is related to the transform, such as the rate of the motion vectors and the rate of the block sizes. In theory, R_p ranges from 0 to $+\infty$.

If $R_p = 0$, no additional bits are spent on the parameter rate, the transform is not performed and the energy is not shifted. If R_p gets larger, more energy will be concentrated to the low band and less is left in the high band. If $R_p \to +\infty$, the clean high band energy can be completely removed and

all the clean signal energy is in the low band. Let $g(R_p)$ be a convex function of R_p indicating the variance of the clean signal in the high band. $g(R_p)$ is a decreasing function, that is, if R_p grows, more energy will be removed from the high band.

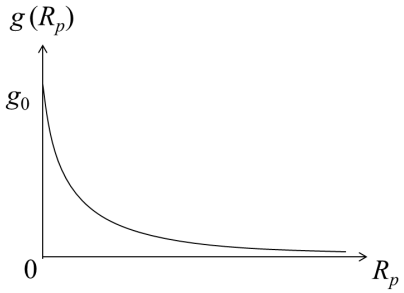

Fig. 2. The variance of the clean high band $g(R_p)$ over R_p.

Let σ_n^2 denote the variance of the noise n_0 or n_1, σ_v^2 the variance of the clean picture, σ_H^2 the variance of the noisy high band, and σ_L^2 the variance of the noisy low band. As the transform is orthonormal, independent noise cannot be shifted. The noisy high band has a variance of

$$\sigma_H^2 = \sigma_n^2 + g(R_p). \tag{1}$$

The variance of the high band is limited to $0 \leq \sigma_H^2 \leq \frac{E}{2}$, where $E = 2\sigma_n^2 + 2\sigma_v^2$ is the total energy. As the transform is orthonormal, the total energy is conserved. The variance of the noisy low band is $\sigma_L^2 = E - \sigma_H^2$.

The theoretical signal model in Fig. 1 helps us to study the relationship between R_p and the entropy of the transform coefficients. To simplify the theoretical model, we assume that the source signal is memoryless Gaussian distributed. Hence, the differential entropies of the temporal bands are

$$h(L) = \frac{1}{2}\log_2(2\pi e) + \frac{1}{2}\log_2(\sigma_L^2), \tag{2}$$

$$h(H) = \frac{1}{2}\log_2(2\pi e) + \frac{1}{2}\log_2(\sigma_H^2). \tag{3}$$

Let h_c be the differential entropy of all coefficients

$$h_c = \frac{1}{2}(h(L) + h(H)) \tag{4}$$

and h_t the total rate of the video signal

$$h_t = R_p + h_c. \tag{5}$$

Note, there is a trade-off between R_p and h_c. Either more bits are spent on R_p to improve the transform while spending less bits for h_c, or less bits are spent on R_p while spending more bits for h_c. For efficient coding, we expect that h_c decreases faster than R_p increases, such that it is possible to reduce the total rate. Theoretically, there exists an optimal R_p^* that minimizes the total rate. Fig. 3 depicts two theoretical graphs of h_t for two different types of orthogonal transforms. If $R_p = 0$, h_t is $\frac{1}{2}(h(L)+h(H))$. If $R_p \to +\infty$, h_c approaches the joint differential entropy $\frac{1}{2}h(\mathbf{x}_1, \mathbf{x}_2)$ of the source signal.

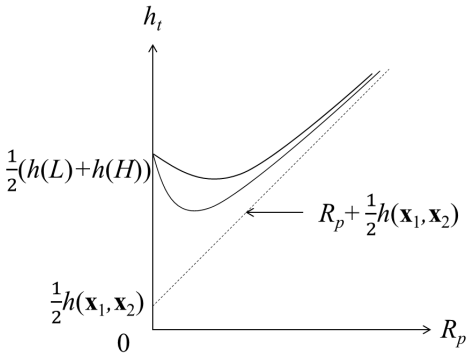

Fig. 3. The total rate h_t over R_p.

We can construct a cost function J_1 from h_c and R_p with $\mu > 0$

$$J_1 = h_c + \mu R_p. \tag{6}$$

With the previous assumptions, we have a convex problem. Hence, we can write the minimization as a constrained problem

$$\min h_c \text{ s.t. } R_p = R_{p_0}, \tag{7}$$

where R_{p_0} is the target rate for the constrained problem. As the variance of the high band is limited to $0 \leq \sigma_H^2 \leq \frac{E}{2}$, h_c is monotonically increasing with σ_H^2. Hence, we simplify the constrained problem

$$\min \sigma_H^2 \text{ s.t. } R_p = R_{p_0}. \tag{8}$$

To solve the simplified constrained problem, we define a new cost function

$$J_2 = \sigma_H^2 + \nu R_p \tag{9}$$

with $\nu > 0$. This cost function can now be used to find the best type of motion-compensated orthogonal transform at each level of a subband decomposition, independent of the quantizer that will be used to quantize the transform coefficients.

To assess the relative performance of one type over another type, the multiplier ν should be larger than zero. This allows us to find the best type of motion-compensated orthogonal transform. Note that only the cost function is used in the following practical coding system, not the theoretical signal model.

III. Practical System

The practical video coding system is depicted in Fig. 4. The input is a group of k pictures (GOP = k). The MCOT is a combination of the unidirectional MCOT [2], the bidirectional MCOT [4], a half-pel motion accurate transform [5], and variable block sizes. The transform is performed as a dyadic decomposition of a group of k frames. After the MCOT, the temporal subbands consist of one temporal low band and $k-1$ temporal high bands. Then the adaptive spatial wavelet transform is applied to the temporal subbands [6]. We apply the spatial decomposition to temporal low band and high bands, because the EBCOT codec requires the same spatial

decomposition level for all the subbands. In this work, the spatial decomposition level is set to three. After the transforms, we use EBCOT to encode the spatial coefficients. According to [7], the uniform deadzone quantization with step size one is used and the rate is controlled by the PCRD.

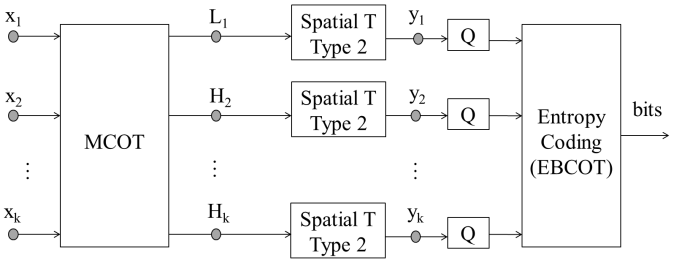

Fig. 4. Video coding system.

A. Construction of Various MCOTs

The purpose of engaging various types of MCOT is that the implemented system is expected to adapt to different video sequences with different contents and patterns and, thus, improve the overall performance. Based on the unidirectional MCOT and the bidirectional MCOT, we are considering the following available transforms to construct the system: intra type, unidirectional MCOT using a past frame, unidirectional MCOT using a future frame, and the bidirectional MCOT.

Intra type means that the original pictures are kept as temporal subbands directly without performing any temporal processing. On a block basis, the intra type is used if the considered motion models fail, that is, blocks that appear only once in a GOP will be intra coded.

Past and future frame unidirectional MCOTs are similar. The only difference is that the past frame unidirectional MCOT considers the temporally previous frame as the reference frame while the future frame unidirectional MCOT considers the temporally subsequent frame.

The bidirectional MCOT takes both previous and subsequent frames into consideration. The system will compare the performance of the four possibilities and choose the most efficient one. The decision is made by minimizing the cost function (9).

B. Obtaining Motion Vectors

Our coding system uses both integer and half-pel accurate motion vectors. To obtain these motion vectors, we minimize the cost function $J_{mv} = \sum |x_i - x_j| + \lambda_{mv} R_{mv}$, where x_i is the reference block and x_j is the current block. $\sum |x_i - x_j|$ is the sum of the absolute differences of the pixel values between the blocks x_i and x_j. λ_{mv} is the Lagrangian multiplier for the motion vectors and R_{mv} is the rate of the motion vectors. Generally, a higher rate for motion vectors can provide a better match between x_i and x_j.

Our motion estimation algorithm provides one integer motion vector (m_{x_0}, m_{y_0}) and its corresponding eight half-pel positions around the integer position per block. However, due to computational complexity, only the integer (m_{x_0}, m_{y_0}) and the best two half-pel motion vectors (m_{x_1}, m_{y_1}) and (m_{x_2}, m_{y_2}) are considered for further evaluation.

Because the motion vectors are crucial to the reconstruction of the video sequences, they require lossless coding. We use Huffman codes for coding.

C. Variable Block Size

In addition to the multiple types of MCOTs, various block sizes are engaged to provide more accurate block-based motion compensation. A macroblock with block size of $m \times n$ is partitioned into smaller block sizes of $m \times \frac{n}{2}$, $\frac{m}{2} \times n$, and $\frac{m}{2} \times \frac{n}{2}$. As discussed above, motion estimation provides a set of motion vectors for each subblock. In our case, there are $1 + 2 + 2 + 4 = 9$ sets of motion vectors saved for each macroblock before the transform. Our system will evaluate all the four subblock partitions to determine which kind of block size is optimal.

In summary, there are three levels that need to be combined inside our system:

- Accuracy of motion compensation for each type of MCOT
- Different types of MCOT for each subblock
- Variable block sizes for each macroblock.

Fig. 5 depicts the structure of the three levels. The first and most detailed level evaluates the nine possible motion vectors for a particular subblock and a particular type of MCOT. The second level evaluates different transform types for each single subblock, given different motion vectors. That is, after the second level, we have a number combinations of motion vectors and transform types for each subblock. Finally, the last level determines which kind of block segmentation is best for a macroblock, given a number of combinations of transforms and motion vectors. In the end, the system gives an efficient combination of motion vectors, transform types, and block partition for each macroblock. Note, the adaptive transforms always allow the open-loop operation of the codec.

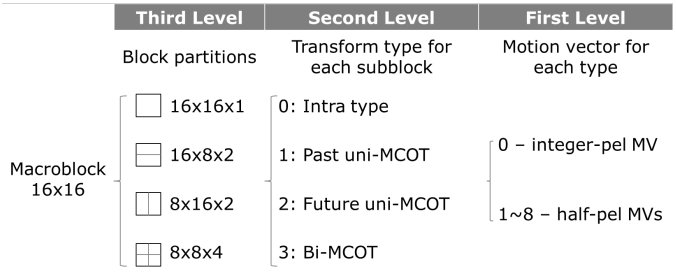

Fig. 5. Structure of the adaptive orthogonal transform.

D. Transform Type Decision

The purpose of our cost function is to determine an efficient combination of motion vectors, transform types, and block partitions. Let R_p be the parameter rate indicating the sum of the rate of the motion vectors $R_p(mv)$, the rate of the types of MCOT $R_p(t)$, and the rate of the block partitions $R_p(b)$. They are obtained from the chosen motion vectors, transform types,

and block partitions, respectively. Let σ_H^2 denote the energy of the high band. We use the cost function (9) in Section II. The Lagrange multiplier ν is chosen to be larger than zero.

The cost function is evaluated on a macroblock level. With all the associated transform types and motion vectors, the transform type which has the smallest cost is chosen for a given multiplier.

IV. Experimental Results

For the experiments, we use the test sequences $Foreman$ and $City$. Motion estimation uses a search range of ± 20. The dictionary for Huffman coding of motion vectors is established from five training sequences, i.e., $Foreman$, $Carphone$, $Salesman$, $Claire$, and $Mother\&Daughter$, each with 288 frames.

Figs. 6 and 7 present the PSNR of the luminance signal over the bit rate for $Foreman$ and $City$ with different transform types. The first graph is the proposed adaptive transform, which is an efficient combination of different transform types, variable block sizes (VBS), and half-pel accurate motion compensation (HP). The second graph is the bidirectional MCOT without being combined with the unidirectional MCOT. The third graph is also the bidirectional MCOT, but without VBS or HP [4]. The fourth graph is the unidirectional MCOT without VBS or HP [2]. The last graph is simply EBCOT-based intra coding for reference.

Fig. 6 demonstrates the advantage over intra coding. The bidirectional MCOT shows a 0.3 dB improvement when compared to the unidirectional MCOT. If variable block size and half-pel motion accuracy is engaged, we have an additional 0.8 dB improvement. Finally, when the adaptive transform permits both unidirectional and bidirectional MCOT, our adaptive transform gains another 0.5 dB when compared to the single bidirectional MCOT. Hence, the adaptive transform always outperforms the individual transforms. The results in Fig. 7 confirm these observations.

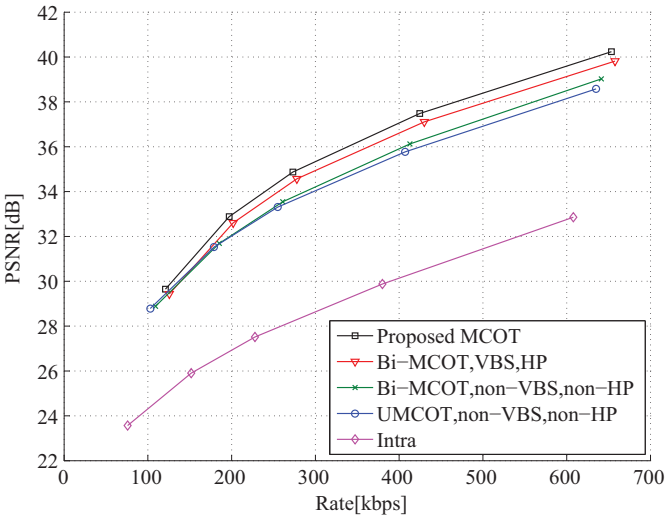

Fig. 6. Luminance PSNR vs. bit rate for the QCIF sequence $Foreman$ at 30fps with 64 frames and a GOP size of 8 frames.

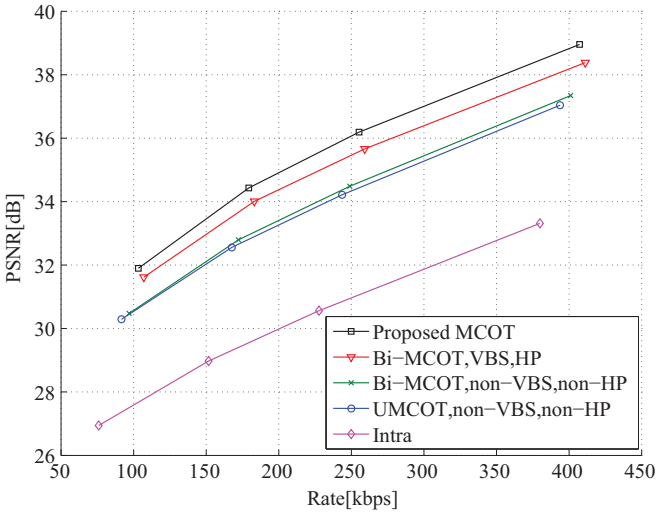

Fig. 7. Luminance PSNR vs. bit rate for the QCIF sequence $City$ at 15fps with 64 frames and a GOP size of 8 frames.

V. Conclusion

This paper discusses a signal model for adaptive motion-compensated orthogonal transforms. A cost function for the adaptive orthogonal transform is constructed to determine the best transform type. This cost function is also used in our practical implementation for transform mode decisions. The second part of the paper describes an efficient combination of several MCOTs. The combination includes multiple types of MCOTs, variable block sizes, and half-pel accurate motion compensation. The experimental results show that our highly adaptive orthogonal transforms improve compression efficiency significantly. The adaptivity is accomplished while maintaining the orthogonality of the overall transform.

References

[1] G. Sullivan and T. Wiegand, "Video compression - from concepts to the H.264/AVC standard," *Proceedings of the IEEE*, vol. 93, no. 1, 2005.

[2] M. Flierl and B. Girod, "A motion-compensated orthogonal transform with energy-concentration constraint," in *Proc. of the IEEE International Workshop on Multimedia Signal Processing*, Oct. 2006, pp. 391–394.

[3] O. Barry, D. Liu, S. Richter, and M. Flierl, "Robust motion-compensated orthogonal video coding using EBCOT," in *Proc. of the Pacific-Rim Symposium on Image and Video Technology*, Singapore, Nov. 2010, pp. 264–269.

[4] M. Flierl and B. Girod, "A new bidirectionally motion-compensated orthogonal transform for video coding," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, Apr. 2007, pp. I–665–I–668.

[5] ——, "Half-pel accurate motion-compensated orthogonal video transforms," in *Proc. of the IEEE Data Compression Conference*, Mar. 2007, pp. 13–22.

[6] M. Flierl, "Adaptive spatial wavelets for motion-compensated orthogonal video transforms," in *Proc. of the IEEE International Conference on Image Processing*, Nov. 2009, pp. 1045–1048.

[7] M. Marcellin, M. Gormish, A. Bilgin, and M. Boliek, "An overview of JPEG-2000," in *Proc. of the IEEE Data Compression Conference*, Mar. 2000, pp. 523–541.