INTER FRAME CODING WITH TEMPLATE MATCHING SPATIO-TEMPORAL PREDICTION

Kazuo Sugimoto, Mitsuru Kobayashi, Yoshinori Suzuki, Sadaatsu Kato, Choong Seng Boon

Multimedia Laboratories, NTT DoCoMo, Inc.
3-5, Hikarinooka, Yokosuka, Kanagawa, Japan

ABSTRACT

A new algorithm is proposed for predicting pixels for inter frame coding without side information. There are many approaches in the past that exploited either spatial or temporal correlations for generating prediction signals of an image in a block-by-block basis. Our method proposed in this paper exploits both spatial and temporal correlations at once to predict the pixels to be encoded. The prediction is achieved by using a template matching mechanism, with reference to previously reconstructed groups of pixels in the same frame or adjacent frames, to fill in the pixels of target regions of a frame. This process is conducted at both the encoder and decoder, and hence allows the decoder to produce the same predictor as the encoder does without any side information. Our coder uses the proposed prediction in addition to conventional motion compensation means .

Simulation results show that our approach achieves up to 11.14% of improvements at the same PSNR over a codec which uses conventional block-based motion compensation.

1. INTRODUCTION

Video coding exploits the implicit correlations present in image sequences in order to achieve tremendous reduction in bitrate. A widely used and well-known approach to reduce the temporal redundancy in image sequences is motion compensation [1] [2]. It is usually performed in two steps: at first, the motion field between a target frame and a reference frame is estimated; then, the motion information is coded and sent to the decoder. At the decoder, the motion information is used to predict the target frame using previously received reference frames. The resulting motion fields, however, are highly correlated and exploitation of this correlation allows greater bitrate reduction. International standard codecs such as H.264[3], employ an approach to reduce the correlation in the motion field domain. Intra prediction [4] is another way to exploit spatial correlation within a frame to reduce the bitrate. But it can not be applied to a block with motion compensation at the same time.

Our approach in this paper, on the other hand, exploits both the spatial and temporal correlation at the same time to predict the pixel values within the target region (block) by applying image filling technique.

An approach which some researchers have tried is texture synthesis as a way to fill in image regions by repeating typical textural patterns. This approach is based on extensive research on texture-synthesis [5][6][7]. Another approach is based on exemplar-based techniques which generate new texture by sampling and copying pixel values from source signals. A number of algorithms are already introduced in image restoration field [8][9]. A. Criminisi, et al. proposed a method of effective image filling algorithm by combining texture-synthesis techniques and exempler-based approach [10].

We have devised an image filling algorithm for inter frame coding to predict the pixel values within a target block. The prediction is achieved by using a template matching mechanism, with reference to previously reconstructed groups of pixels in the same frame or adjacent frames, to fill in the pixels of target regions of a frame. This process is conducted at both the encoder and decoder, and hence allows the decoder to produce the same predictor as the encoder does without any side information. We call this prediction technique as Template Matching Spatio-Temporal (TMST) prediction.

In section 2, we present the proposed Template Matching Spatio-Temporal prediction technique. Section 3 describes the encoder and decoder algorithm using the proposed prediction technique. Simulation results are shown in section 4 followed by conclusion in section 5.

2. PREDICTION BY SPATIO-TEMPORAL TEMPLATE MATCHING TECHNIQUE

The proposed prediction algorithm is based on the exemplar-based image filling technique introduced by A. Criminisi, et al. Given a target block of a frame, a target pixel in the block is determined by finding an optimum pixel from a set of reference samples, where the adjacent pixels of the optimum pixel have the highest correlation with those of the target pixel. Here, the adjacent pixels of the target pixel are assumed to be available through previous prediction process. The reference samples may

0-7803-8554-3/04/$20.00 ©2004 IEEE.

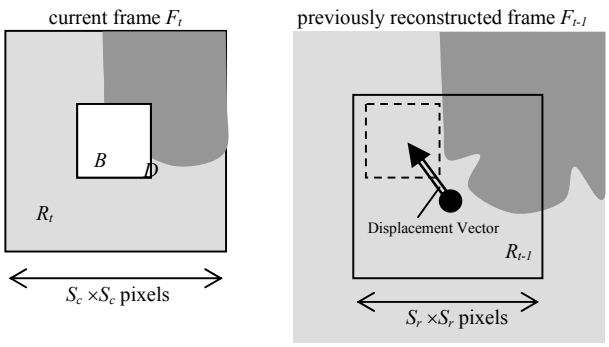

Figure 1: Target region B and reference region R

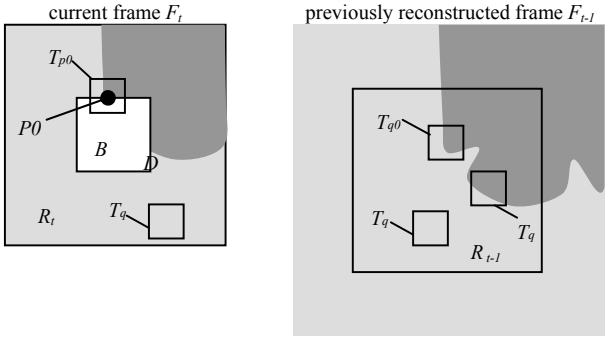

Figure 2: Template matching

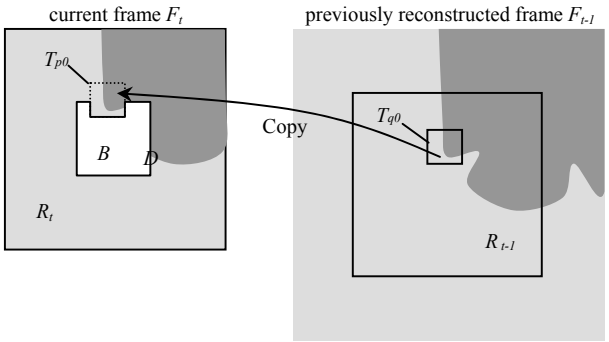

Figure 3: Replication

come from either the same frame or a different frame, and hence the pixels within a target block are results of intra and inter prediction. This is distinctively different from the conventional prediction methods.

For clarity, the following expressions are defined:

- **Target block:** a block which is to be filled in
- **Readily-available-pixels:** pixels which are produced through previously reconstructing or filling in process
- **Reference region:** a region which contains the pixels which are readily available

As shown in figure 1, a target block to be filled in is indicated by B. The pixels at the boundary of B are indicated by D. The boundary pixels D evolve towards the inside of B as the algorithm progresses. The reference region, indicated by R, contains the pixels which are readily available through previous reconstruction (R_{t-1}) or filling in process (R_t). R encompasses reconstructed pixels in both the current frame (F_t) to be reconstructed and the previously reconstructed frame (F_{t-1}).

In the current frame F_t, the reference region R_t is a square region of $S_c \times S_c$ pixels, centered at the target block B. In the previously reconstructed frame F_{t-1}, the reference region R_{t-1} is a square region of $S_r \times S_r$ pixels centered at a block pointed by a displacement vector. The displacement vector will be described in the next section. Now, consider a block in frame F_t that straddles the boundary between a target block B and a region containing readily-available-pixels. This region, hereafter known as target template T_p is centered at a pixel p, where p belongs to D. The best-matched template of T_p is searched within the reference region R. The optimum template, T_{q0}, gives the highest correlation among the readily-available-pixels of T_p and the corresponding pixels in T_{q0}.

In the proposed algorithm, first, the order in which the target pixels to be filled is determined, then the value of the pixels are determined by the template matching mechanism described above. The both steps iterate until all pixels in the target have been filled.

1. Filling order determination

Filling order is crucial in non-parametric texture synthesis. The priority of the filling process is biased toward those pixels in the target template which lie along strong edges and which are surrounded by larger number of readily-available-pixels.

Given a pixel $p \in dB$, its priority $P(p)$ is defined as follows:

$$P(p) = N(p) \cdot E(p) \quad (1)$$

where $N(p)$ indicates the number of readily-available-pixels around the eight neighboring positions of pixel p.

$E(p)$ is defined as follows:

$$E(p) = |\nabla I_p \cdot l_p| \quad (2)$$

where ∇I_p is the gradient of the pixels adjacent to the pixel p, and l_p is a unit vector along the contour D at the pixel p. After computing all the priorities of the pixels on the contour D, the target template T_p whose center pixel $p0$ provides the highest priority as depicted in figure 2 is selected for the replication step.

2. Template matching and replication

Once the highest priority template T_{p0} is determined, template matching to find the optimum reference template is performed. Next, replication step is performed by copying the pixels which are not filled in within the target template. The undefined pixels in the target template are filled with the corresponding pixels in the optimum reference template $T_{q0} \in R$.

Formally,

$$T_{q0} = \arg\min_{T_q \in R} d(T_{p0}, T_q) \quad (3)$$

where the distance $d(T_a, T_b)$ between two templates, T_a and T_b is simply defined as the sum of squared differences of the readily-available-pixels in the two templates. Having found the reference template T_{q0} by template matching as in figure 2, the value of each pixel to be filled, $Q' | Q' \in T_{p0} \cap B$, is copied from its corresponding position inside T_{q0} as in figure 3.

After the replication step, the pixels which are filled in are assigned to the reference region R.

3. ENCODER AND DECODER ALGORITHM

We apply the combination of conventional motion compensation and the proposed prediction method for inter frame encoding/decoding of a block based codec architecture. Blocks encoded with conventional motion compensation are called M-mode blocks, and those with the TMST prediction are called P-blocks. Then the prediction residuals are coded with DCT and quantization followed by entropy coding.

For the proposed prediction method, the template matching, as described above, is performed in integer-pel precision when referenced to the current frame and in half-pel precision when referenced to the previously reconstructed frame.

We used YUV4:2:0 color space and performed the TMST prediction on Y component. Pixels of U and V components are filled by copying the pixel values at the corresponding positions of the Y component.

ENCODER

The algorithm of the encoder is as below:

1. Mode selection
2. Coding/Reconstruction of M-mode blocks
3. TMST prediction for P-mode blocks
4. Coding/Reconstruction of P-mode blocks
5. Entropy Coding

Mode selection is performed as follows. First, motion compensated frame is constructed using conventional motion estimation compensation techniques. Then for each block, the efficiency of the TMST prediction is tested by comparing to that of the conventional technique. If the prediction error using the TMST prediction is smaller, then P-mode is assigned as the prediction mode for that block, otherwise M-mode is assigned.

After the prediction modes of all blocks are determined, the residuals of the blocks are coded using DCT / quantization. For reference of subsequent prediction, the coded frames are reconstructed. Motion vectors are predicted by median prediction like MPEG-4,

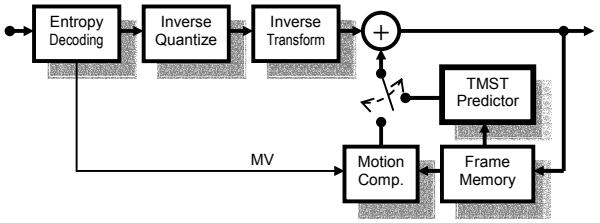

Figure 4: Diagram of the decoder

and the prediction residuals of motion vectors are coded. Motion vector prediction is performed also for each P-block to define the displacement vector used in the template matching process. The same motion vector is used for motion vector prediction of other blocks.

Note that the template matching process will also be conducted at the decoder, as described below. Therefore, unlike the conventional motion compensation technique, the encoder needs not send any side information for prediction of P-blocks, and thus allows further reduction of bitrates.

One of the most challenging problems in this coding algorithm is to determine the prediction mode for each block. The pixels in P-mode blocks which reside subsequent to a target P-mode block, in raster scan order, can not be used as the reference, for they are not reconstructed before coding of the target block. This implies that the existence of a subsequent P-mode block, in raster scan order, affects the prediction of the target P-mode block. Here we pre-process all the blocks with M-mode first, and then apply the proposed prediction for each block later in reverse raster scan order. By this, the ambiguity when performing mode selection is reduced.

DECODER

The block diagram of the decoder is shown in figure 4. As shown in figure 4, the residual signals are reconstructed through entropy decoding, inverse quantization, and inverse transform. Then all the M-mode blocks are decoded. Subsequently, all P-mode blocks are decoded. Because the proposed prediction uses only reconstructed pixels as the reference, decoder can perform exactly the same prediction as at the encoder without the additional side information.

4. SIMULATION AND RESULTS

To see the effectiveness of the proposed coding algorithm, we have performed a simulation using “Foreman” sequence at 7.5fps, QCIF resolution. With the proposed algorithm, an image is divided into blocks of 8×8 pixels. The result is compared to a block-based motion compensation codec as in MPEG-4.

Table 1 compares the results of the codecs and Figure 5 shows their R-D curves. It can be observed that the proposed algorithm outperforms the conventional prediction up to 11.14% at the same PSNR. Figure 6 shows the blocks in a frame which are encoded as P-mode blocks while Figure 7 shows an example of how a block is built up using the proposed algorithm.

As we exploited both spatial and temporal correlation of an image in predicting a P-mode block, the proposed scheme can perform prediction in a more flexible manner than the conventional block-based prediction scheme.

5. CONCLUSIONS AND FUTURE WORK

We proposed an efficient prediction method for inter frame coding which exploits both spatial and temporal correlation in an image. We applied the prediction to a coder, and proved that our method works better than the conventional one up to 11.14% in terms of bitrate at the same PSNR. Mode selection is one of the key techniques which should be further investigated, and the prediction algorithm will also be investigated further in the future.

6. REFERENCES

[1] J. R. Jain and A. K. Jain, "Displacement Measurement and its Application in Interframe Image Coding", IEEE Transactions on Communications, Vol. 29, No. 12, pp. 1799-1808, Dec. 1981.

[2] F. Dufaux and F. Moscheni, "Motion Estimation Techniques for Digital TV: A Review and a New Contribution", Proceedings of the IEEE Vol. 83, No. 6, pp.858-876, June 1995.

[3] Joint Vide Team (JVT), "Study of Final Committee Draft of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC)", Document JVT-F100d2, Generated on December 2002.

[4] B. Meng, Au O. C. , C. Wong, H. Larn, "Efficient intra-prediction algorithm in H.264", ICIP 2003, Vol. 3, pp. 837-840, Sept. 2003.

[5] D. Garber, "Computational Models for Texture Analysis and Texture Synthesis", Ph.D thesis, Univ. of Southern California, USA, 1981.

[6] R. Bornard, E. Lecan, L. Laborelli, and J-H. Chenot, "Missing data correction in still images and image sequences", ACM Multimedia, France, Dec. 2002.

[7] M. Ahikhmin, "Synthesizing natural textures", Proc. ACM Symp. On Interactive 3D Graphics, pp. 217-226, Research Triangle Park, NC, Mar. 2001.

[8] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image Inpainting", Proc. ACM Conf. Comp. Graphics(SIGGRAPH), pp.417-424, New Orleans, LU, Jul. 2000.

[9] T. F. Chan and J. Shen, "Non-texture inpainting by curvature-driven diffusions(CDD)", J. Visual Comm. Image Rep., 4(12), 2001

[10] A. Criminisi, P. Perez, and K. Toyama, "Object removal by exemplar-based inpainting", Proc. Conf. Comp. Vision Pattern Rec., Madison, WI, Jun 2003.

Table 1: Comparison of the conventional approach and the proposed method.

conventional		proposed			improved
PSNR [dB]	Bitrate [kbps]	PSNR [dB]	Bitrate [kbps]	Selected block[%]	Bitrate [%]
38.32	141.6	38.41	135.1	23.74	-4.58
35.39	94.2	35.48	90.3	16.41	-4.20
31.88	52.7	31.92	49.5	14.90	-6.04
28.47	32.7	28.49	29.8	21.97	-8.92
25.42	17.2	25.48	15.3	14.65	-11.14

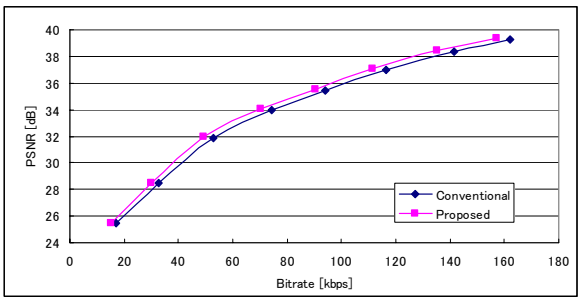

Figure 5: R-D curves of the conventional approach and the proposed method.

Figure 6: An example of prediction mode selection; Black squares are P-mode blocks.

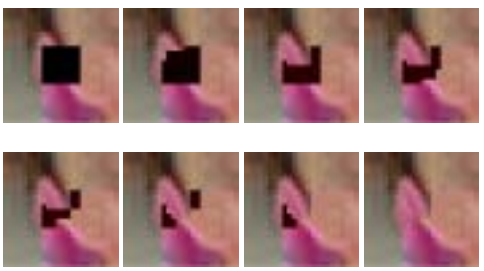

Figure 7: Evolution of filling up a P-mode block.