14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

LOSSLESS VIDEO CODING USING MULTI-FRAME MC AND 3D BI-PREDICTION OPTIMIZED FOR EACH FRAME

Hiroki Maeda, Akira Minezawa, Ichiro Matsuda and Susumu Itoh

Department of Electrical Engineering, Faculty of Science and Technology,
Science University of Tokyo
2641 Yamazaki, Noda-shi, Chiba 278-8510, JAPAN
Tel: +81 4 7124 1501 (ext.3722); Fax: +81 4 7124 9367
E-mail: maeda@itohws01.ee.noda.tus.ac.jp

ABSTRACT

This paper proposes an efficient lossless video coding scheme based on forward-only 3D bi-prediction. In this scheme, a video signal at each pel is predicted using not only the current frame but also two motion-compensated reference frames. Since both the reference frames are taken from the past, coding process of successive frames can be performed in temporal order without extra coding delay. The resulting prediction errors are encoded using context-adaptive arithmetic coding. Several coding parameters, such as prediction coefficients and motion vectors, are iteratively optimized for each frame so that an overall coding rate required for the frame can be a minimum. Experimental results indicate that coding rates of the proposed scheme are 14–21 % lower than those of the H.264/AVC-based lossless coding scheme.

1. INTRODUCTION

Motion compensation (MC) is an essential component for lossy video coding and several improved techniques, such as bi-directional prediction and variable block-size MC, have been developed. Nevertheless, it is only recently that MC techniques have been used for lossless video coding [1, 2]. One of most promising ways of incorporating the MC technique into lossless video coding is an MC based 3D prediction method [1]. In this method, a 3D linear predictor, which predicts a video signal using both the current frame and a motion compensated reference frame, is optimized at each pel. However, such a pel-by-pel optimization increases complexity of not only encoding but also decoding processes. To cope with this problem, we proposed an efficient lossless video coding scheme based on variable block-size MC and block-adaptive 3D prediction [3]. Since the scheme optimizes a set of 3D predictors as well as motion vectors only at the encoder side and transmits them to the decoder as side information, it can be reasonably fast in decoding process. This nature is suitable for video applications where decoding should be always in real-time.

In typical lossy video coding standards [4, 5], bi-directional prediction based on a linear combination of forward and backward prediction is often employed to improve coding efficiency [6]. Motivated by this fact, we have examined bi-directional 3D prediction which uses the previous and following frames as the reference frames [7]. As a result, coding rates of B-frames where the bi-directional 3D prediction is applied are considerably reduced. However, since periodical insertion of B-frames increases temporal distance between the current and reference frames, coding efficiency of the remaining P-frames becomes worse. Furthermore, intentional degradation of B-frame quality with a slightly coarser quantization parameter, which usually improves the overall coding performance in terms of a rate-distortion sense, is not allowed in lossless coding. Consequently, a total coding gain obtained by using the bi-directional 3D prediction is rather limited in our experiments.

On the other hand, the H.264/AVC [5], which is the latest lossy video coding standard, introduces a new prediction method called bi-prediction for generalized B-frames [8]. The bi-prediction uses two reference frames like the conventional bi-directional prediction, however both reference frames can be freely selected from already encoded frames. It allows forward-only bi-prediction without extra coding delay and all frames except for the first and second frames in a group of pictures (GOP) can take advantage of the generalized B-frames. In addition, because of its simple GOP structure, the forward-only bi-prediction is suitable for incorporating with a multi-frame MC technique which adaptively selects the reference frames in each block.

In this paper, we extend our lossless coding scheme to allow an H.264/AVC-like forward-only bi-prediction method and evaluate its coding performance through some experiments.

2. MC BASED 3D BI-PREDICTION

Like the existing lossy video coding standards [4, 5], the proposed coding scheme segments a video sequence into a number of GOPs which consist of the three frame types called I-, P- and B-frames. In this paper, the first and second frames in a GOP are encoded as I- and P-frames respectively, and the remaining frames are all treated as B-frames. Figure 1 illustrates the proposed 3D bi-prediction method used in B-frames. In the figure, $\boldsymbol{p}_0$ is a pel to be predicted and $\boldsymbol{p}_k$s ($k = 1, 2, \ldots, 20$) are already encoded pels in the current frame. In addition, two groups of pels $\boldsymbol{q}_k$s ($k = 1, 2, \ldots, 25$) and $\boldsymbol{q}'_k$s ($k = 1, 2, \ldots, 13$), which belong to the primary and secondary reference frames, are also used for the prediction. To exploit temporal correlations effectively, both groups of pels are motion-compensated, that is their positions are shifted according to motion vectors detected for the respective reference frames. This pair of the forward motion vectors $\boldsymbol{v}$ and $\boldsymbol{v}'$ is given in each square block of variable-size. The block is called an MC-block and its size is adapted using quadtree partitioning [3]. Furthermore, the current frame is uniformly divided into small blocks composed of 8×8 pels and each block is classified into one of 24 classes ($m = 1, 2, \ldots, 24$). Each class has an individual predictor

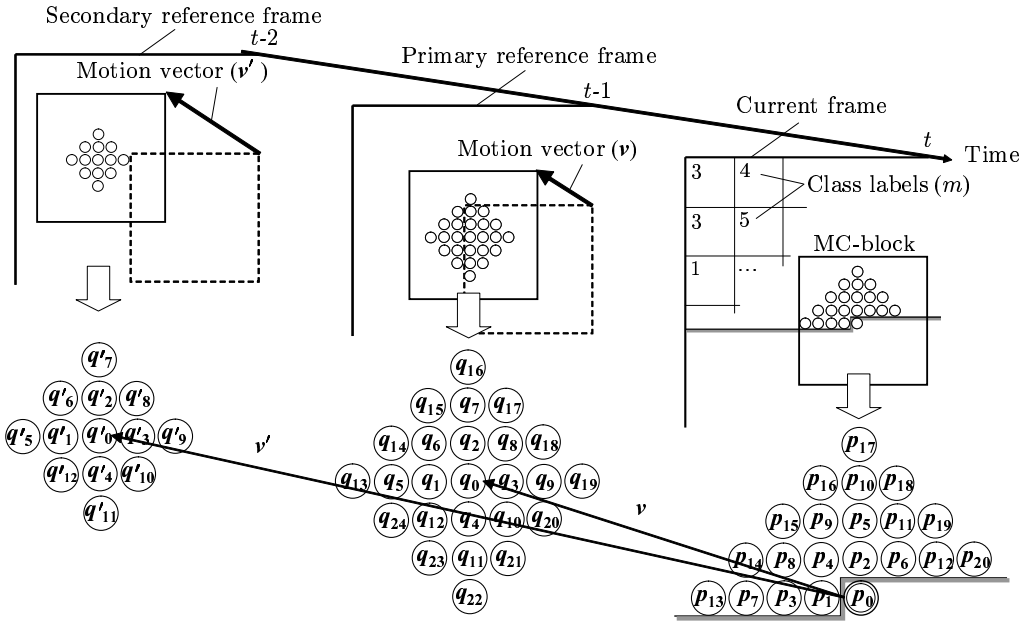

Figure 1: Block-adaptive 3D bi-prediction.

which is optimized for blocks belonging to the same class.

This kind of classification-based block-adaptive prediction [9] is carried out for all of the three frame types in a similar way. In general form, therefore, a predicted value $\hat{s}(\boldsymbol{p}_0)$ is expressed as:

$$\begin{aligned}\hat{s}(\boldsymbol{p}_0) = &\sum_{k=1}^{K_p} a_m(k)\cdot s(\boldsymbol{p}_k) \\ &+ \sum_{k=0}^{K_q-1} a_m(k+K_p+1)\cdot s(\boldsymbol{q}_k) \\ &+ \sum_{k=0}^{K'_q-1} a_m(k+K_p+K_q+1)\cdot s(\boldsymbol{q}'_k), \end{aligned} \tag{1}$$

where $a_m(k)$s ($k = 1,2,\ldots,K_p+K_q+K'_q$) are prediction coefficients of the m-th predictor and $s(\boldsymbol{p}_k)$ represents a value of a video signal at a pel $\boldsymbol{p}_k$. K_p, K_q and K'_q are the numbers of pels taken from the current and the two reference frames for the prediction. By ignoring the second and third terms in the right hand side of Eq.(1), it represents 2D prediction used in an I-frame. In this case, no motion vector is required. For a P-frame, only the third term and the motion vector $\boldsymbol{v}'$ which concern the secondary reference frame are omitted. In general, use of higher prediction order, that is a larger value of $K_p+K_q+K'_q$, improves prediction accuracy and yields a lower coding rate of prediction errors. However, it increases the amount of side information on prediction coefficients $a_m(k)$s at the same time. From this point of view, we tested several combinations of the parameters K_p, K_q and K'_q for each frame type, and found that values listed in Table 1 are reasonable in terms of the overall coding performance. Accordingly, these values are used in the rest of the paper.

Table 1: The number of pels used for the prediction.

Frame type	K_p	K_q	K'_q	Total
I-frame	30	–	–	30
P-frame	20	25	–	45
B-frame (forward-only)	20	25	13	58

3. CODING OF PREDICTION ERRORS

After the prediction, context modeling for adaptive arithmetic coding of a prediction error $e = s(\boldsymbol{p}_0) - \hat{s}(\boldsymbol{p}_0)$ is performed at each pel $\boldsymbol{p}_0$. The context modeling is based on non-linear quantization of a context function $U(\boldsymbol{p}_0)$ which captures statistical property of the prediction errors in neighboring areas of the pel $\boldsymbol{p}_0$. In this paper, the context function is defined as the sum of absolute prediction errors at already encoded pels [3]. For example, the context function used for B-frames is given by:

$$\begin{aligned}U(\boldsymbol{p}_0) = &\sum_{k=1}^{6} \left| s(\boldsymbol{p}_k) - \hat{s}(\boldsymbol{p}_k) \right| \\ &+ \sum_{k=0}^{4} \left\{ \left| s(\boldsymbol{q}_k) - \hat{s}(\boldsymbol{q}_k) \right| + \left| s(\boldsymbol{q}'_k) - \hat{s}(\boldsymbol{q}'_k) \right| \right\}. \end{aligned} \tag{2}$$

Each quantization level of $U(\boldsymbol{p}_0)$ corresponds to one of sixteen contexts ($n = 1,2,\ldots,16$) and thresholds $\{Th_m(1), Th_m(2), \ldots, Th_m(15)\}$ used in this quantization are optimized for each class (m) as described later. In consequence of this context modeling, we assume that a conditional probability density function (PDF) of the prediction error e observed in each context can be modeled by the generalized Gaussian function [9]:

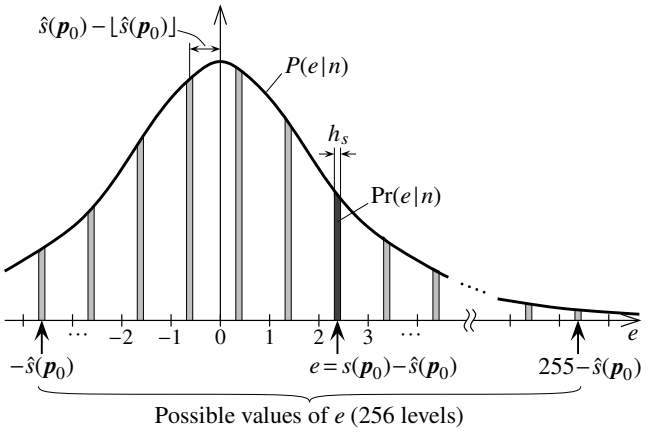

Figure 2: Conditional probability of occurrence of the prediction error e.

$$P(e|n) = \frac{c_n \cdot \eta(c_n, \sigma_n)}{2\Gamma(1/c_n)} \cdot \exp\left\{-\left|\eta(c_n, \sigma_n) \cdot e\right|^{c_n}\right\},$$

$$\eta(c_n, \sigma_n) = \frac{1}{\sigma_n} \sqrt{\frac{\Gamma(3/c_n)}{\Gamma(1/c_n)}}, \tag{3}$$

where $\Gamma(\cdot)$ is the gamma function, σ_n is a standard deviation of e and c_n is a shape parameter which controls sharpness of the PDF. Since 8-bit monochrome video signals are expressed as integer values from 0 to 255, possible values of the prediction error e for a given $\hat{s}(\boldsymbol{p}_0)$ are also limited to the following 256 values:

$$e \in \{s - \hat{s}(\boldsymbol{p}_0) \mid s = 0, 1, 2, \ldots, 255\}. \tag{4}$$

Therefore, a conditional probability of occurrence for each possible value of e, when both the context n and the predicted value $\hat{s}(\boldsymbol{p}_0)$ are known, is derived from the above PDF.

$$\Pr(e|\hat{s}(\boldsymbol{p}_0), n) = \frac{\Pr(e|n)}{\sum_{s=0}^{255} \Pr(s - \hat{s}(\boldsymbol{p}_0)|n)}, \tag{5}$$

$$\Pr(e|n) = \int_{-h_s/2}^{h_s/2} P(e + \varepsilon|n)\, \mathrm{d}\varepsilon. \tag{6}$$

As a matter of fact, the predicted value $\hat{s}(\boldsymbol{p}_0)$ is explicitly rounded to the nearest multiple of $h_s = 1/8$ to avoid accumulation of unexpected rounding errors. Hence the value of h_s is used as an interval for integration of the PDF in Eq.(6). Adaptive arithmetic coding of the actual value of e is carried out according to the conditional probabilities calculated by using Eqs.(5) and (6). Note that the numerator of Eq.(5) corresponds to the area shown in dark gray and the denominator is the sum of the shaded areas in Figure 2. Practically, by storing all of the probabilities in a look-up table at a sampling rate of $1/h_s$, we can considerably reduce computation required for the adaptive arithmetic coding.

4. OPTIMIZATION OF CODING PARAMETERS

In the proposed lossless video coding scheme, parameters listed below are optimized for each frame and transmitted to the decoder as side information [7].

- Quadtree for MC-block size.
- Motion vectors $\boldsymbol{v}$ and $\boldsymbol{v}'$ for each MC-block.
- Class label m for each block of 8×8 pels.
- Prediction coefficients $a_m(k)$s for each class.
- Thresholds $\{Th_m(n)\}$ for each class.
- Shape parameter c_n for each context.

Optimization of these coding parameters is done by iteratively minimizing the following cost function:

$$J = -\sum_{\boldsymbol{p}_0} \log_2 \Pr(e|\hat{s}(\boldsymbol{p}_0), n) + B_{side}. \tag{7}$$

The first term of the cost function represents the number of coding bits required for the prediction errors. The second term (B_{side}) is the amount of side information on the above coding parameters. Concrete procedures for the optimization in a B-frame, for example, are as follows.

(1) Initial motion vectors $\boldsymbol{v}$ and $\boldsymbol{v}'$ are determined in each MC-block composed of 16×16 pels by using the block matching algorithm.
(2) Provisional classification of the small blocks composed of 8×8 pels is carried out and initial predictors are designed for the respective classes.
(3) Partial optimization of the predictor is performed by gradually varying two prediction coefficients $a_m(i)$ and $a_m(j)$ which are chosen randomly. This operation is repeated a certain number of times for every class.
(4) The thresholds $\{Th_m(1), Th_m(2), \ldots, Th_m(15)\}$ are optimized in each class by using the dynamic programming technique.
(5) The optimum value of the shape parameter c_n is selected in each context.
(6) All the predictors are tested for each small block and the optimum predictor, or the optimum class is selected for the block.
(7) Refinement of motion vectors $\boldsymbol{v}$ and $\boldsymbol{v}'$ are performed within quadtree-based three-level decomposition of MC-blocks from 32×32 to 8×8 pels. As a result, the best combination of the MC-block size and the motion vectors are determined.
(8) The above procedures (3)–(7) are repeated until all the coding parameters converge in each frame.

5. REFERENCE FRAME SELECTION FOR 3D BI-PREDICTION

Multi-frame MC is a technique to improve prediction accuracy by using multiple reference frames. In the bi-prediction method of the H.264/AVC standard [5], up to two reference frames are selected from among a certain number of decoded frames block-by-block. These reference frames are motion compensated and then blended to obtain predicted values for the block. Blending ratio of the two reference frames is changeable when a coding tool of weighted prediction is enabled. In this paper, a similar technique is introduced into the proposed 3D bi-prediction method. Specifically, past decoded R frames indexed by time instants $t-1, t-2, \ldots, t-R$ are stored in memory as candidates for the reference frames. Since the most recent decoded frame $(t-1)$ is very likely to contain useful information for the prediction, it is always used as the primary reference frame and only the secondary reference frame is selected from the

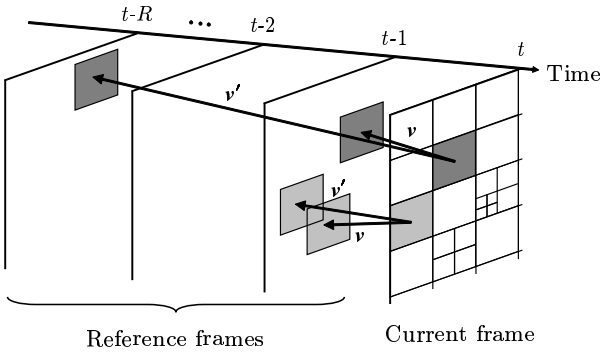

Figure 3: Multi-frame MC in the proposed scheme.

R candidates in each MC-block as shown in Figure 3. This restriction greatly reduces complexity of the encoder and can even improve coding performance slightly because the side information specifying the primary reference frame is not needed. The selection of the secondary reference frame is carried out so that the cost-function J can be a minimum in the optimization process (7) described in Section 4.

Incidentally, recent lossy video coding schemes commonly employ the MC technique based on motion vectors with fractional-pel accuracy [10]. On the other hand, we restrict both motion vectors $\boldsymbol{v}$ and $\boldsymbol{v}'$ to integer-pel accuracy. However, in our scheme, the predicted value includes the weighted sum of several pels in the reference frame as shown in Eq.(1). It means the proposed scheme can conduct interpolation of spatially adjacent pels, which is needed for the conventional MC with fractional-pel accuracy, in a more flexible way. Besides, the weighted prediction based on adaptive blending of the two reference frames is also realized within a framework of the proposed scheme.

6. EXPERIMENTAL RESULTS

To evaluate coding performance of the proposed scheme, some experiments are conducted using CIF-sized (352×288 pels) monochrome video sequences shown in Figure 4. In the experiments, the first 25 frames of each sequence are encoded as a single GOP. The range coder [11], which is known as a fast implementation of a multi-symbol arithmetic coder, is used for entropy coding of prediction errors and side information.

Figure 4: Test sequences.

Figure 5 indicates bit-rate savings obtained by using the forward-only 3D bi-prediction. In this figure, $R = 1$ means the reference scheme [3] where no B-frame is used (i.e. IPPPP$\cdots$). Meanwhile, a GOP structure of the proposed scheme ($R \geq 2$) is always IP$BBB\cdots$, where an italic letter B means the generalized B-frame based on the forward-only bi-prediction. It is shown that the proposed scheme obviously outperforms the reference scheme ($R = 1$) owing to use of two reference frames. The coding gains generally increase as the number of reference frames (R), however they are quickly saturated when R is larger than 5. Coding rates of the 'Tempete' sequence measured at each frame are shown

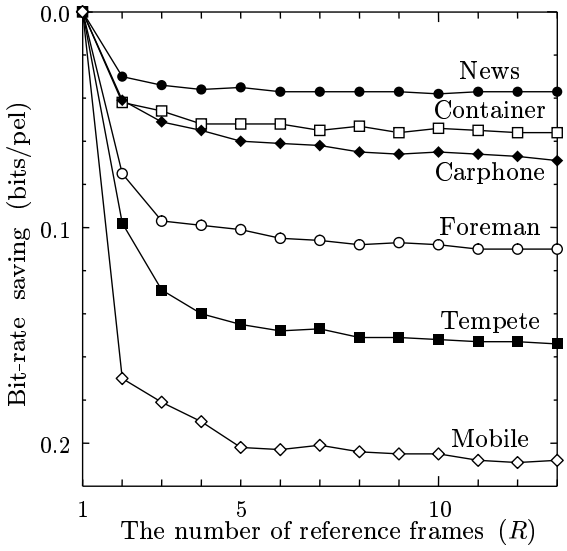

Figure 5: Bit-rate savings obtained by forward-only 3D bi-prediction.

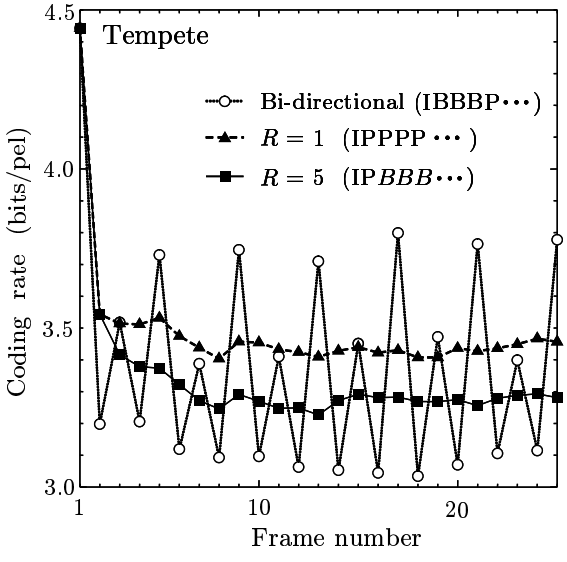

Figure 6: Coding rates of the 'Tempete' sequence.

Table 2: Comparison of coding rates (bits/pel).

Sequence	$R = 5$ (IP*BBB*$\cdots$)	$R = 2$ (IP*BBB*$\cdots$)	Bi-directional [7] (IBBBP$\cdots$)	FRExt [12] (IP*BBB*$\cdots$)	JPEG-LS [13] (I I I I I$\cdots$)
Carphone	**2.651**	2.669	2.690	3.089	3.493
Container	2.274	2.283	**2.268**	2.785	4.144
Foreman	**2.719**	2.746	2.786	3.178	3.881
Mobile	3.474	3.506	**3.469**	4.168	5.403
News	**1.312**	1.318	1.331	1.656	3.412
Tempete	**3.345**	3.392	3.392	3.958	4.848
Average	***2.629***	*2.652*	*2.656*	*3.139*	*4.197*

in Figure 6. In the figure, 'Bi-directional' indicates the previously reported scheme based on bi-directional 3D prediction where the previous and following frames are used as the primary and secondary reference frames respectively [7]. The numbers of pels used for the bi-directional 3D prediction are set to $K_p = 20$ and $K_q = K'_q = 25$, and three successive B-frames and a P-frame are periodically assigned in a GOP (i.e. IBBBP$\cdots$). These conditions were found to give the best results in our preliminary experiments. The coding rate of 'Bi-directional' drastically fluctuates depending on frame types in the GOP, and inefficient results at P-frames make the overall coding performance decrease. On the other hand, the proposed scheme ($R = 5$) shows stable coding performance and improvement from the reference scheme ($R = 1$) is constantly obtained at every *B*-frame.

Finally, Table 2 compares coding rates of some lossless coding schemes. 'FRExt' means a lossless mode of the Fidelity Range Extension of H.264/AVC [12]. We employ the FRExt reference software JM 10.1 with the conditions of Context-based Adaptive Binary Arithmetic Coding (CABAC) and multi-frame MC with 5 reference frames. 'JPEG-LS' is an intra-frame coding scheme which utilizes the JPEG-LS algorithm [13] on a frame-by-frame basis. The coding rate shown in **boldface** represents the best result for each sequence. We can see that coding performance of 'Bi-directional' and the proposed scheme of $R = 2$, both of which use two reference frames in the 3D prediction, are almost the same. However, the proposed scheme can be improved by introducing the multi-frame MC technique and its average coding rate is 0.027 bits/pel lower than 'Bi-directional' when the number of the reference frames is $R = 5$.

7. CONCLUSIONS

We have proposed an efficient lossless video coding scheme based on forward-only 3D bi-prediction. The scheme is similar to our previously reported scheme based on bi-directional 3D prediction [7] in the sense that both schemes use two motion-compensated reference frames in the prediction. Experimental results show that coding performance of the proposed scheme is almost the same as the previous scheme when two reference frames are used. However, the proposed scheme has an advantage in that the performance can be easily improved by introducing the multi-frame MC technique. As a result, the proposed scheme attains 14–21 % better coding performance than the H.264/AVC-based lossless coding scheme.

REFERENCES

[1] D. Brunello, G. Calvagno, G. A. Mian and R. Rinaldo, "Lossless Compression of Video Using Temporal Information," IEEE Trans. on Image Processing, Vol. 12, No. 2, pp. 132–139, Feb. 2003.

[2] S. Sun and S. Lei, "On Study of a Lossless Video Coding Algorithm Based on H.26L Tools," Proc. SPIE Conf. on Image and Video Communications and Processing (IVCP-2003), Vol. 5022, pp. 994–1000, Jan. 2003.

[3] I. Matsuda, T. Shiodera and S. Itoh, "Lossless Video Coding Using Variable Block-Size MC and 3D Prediction Optimized for Each Frame," Proc. 12th European Signal Processing Conf. (EUSIPCO 2004), pp. 1967–1970, Sep. 2004.

[4] ITU-T Rec. H.262 | ISO/IEC 13818-2, "Information Technology – Generic Coding of Moving Pictures and Associated Audio Information: Video," 1995.

[5] ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC, "Advanced Video Coding for Generic Audiovisual Services," 2003.

[6] B. Girod, "Why B-Pictures Work: A Theory of Multi-Hypothesis Motion-Compensated Prediction," Proc. 1998 Int. Conf. on Image Processing (ICIP '98), Vol. II, pp. 213–217, Oct. 1998.

[7] I. Matsuda, T. Shiodera, H. Maeda and S. Itoh, "Lossless Video Coding Using Bi-Directional 3D Prediction Optimized for Each Frame," Proc. European Conf. on Circuit Theory and Design (ECCTD 2005), Vol. II, pp. 71–74, Sep. 2005.

[8] M. Flierl and B. Girod, "Generalized B Pictures and the Draft H.264/AVC Video-Compression Standard," IEEE Trans. on Circuits and Systems for Video Technology, Vol. 13, No. 7, pp. 587–597, July 2003.

[9] I. Matsuda, N. Shirai and S. Itoh, "Lossless Coding Using Predictors and Arithmetic Code Optimized for Each Image," Proc. 8th Int. Workshop VLBV03, Lecture Notes in Computer Science Vol. 2849, pp. 199–207, Sep. 2003.

[10] B. Girod, "Motion-Compensating Prediction with Fractional-Pel Accuracy," IEEE Trans. on Communications, Vol. 41, No. 4, pp. 604–612, Apr. 1993.

[11] M. Schindler, "A Fast Renormalization for Arithmetic Coding," Proc. IEEE Data Compression Conf. (DCC '98), p. 572, 1998.

[12] G. J. Sullivan, P. Topiwala and A. Luthra, "The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extensions," Proc. SPIE Conf. Applications of Digital Image Processing XXVII, Vol. 5558, pp. 53–74, Aug. 2004.

[13] ISO/IEC, ISO/IEC 14495-1:1999, "Information Technology – Lossless and Near-lossless Compression of Continuous-Tone Still Images: Baseline," Dec. 1999.