BLOCK-BASED SPATIO-TEMPORAL PREDICTION FOR VIDEO CODING

Ichiro Matsuda, Kyohei Unno, Hisashi Aomori and Susumu Itoh

Department of Electrical Engineering, Faculty of Science and Technology,
Tokyo University of Science

2641 Yamazaki, Noda-shi, Chiba 278-8510, JAPAN
phone: +81 4 7124 1501 (ext.3740), fax: +81 4 7124 9367
email: matsuda@ee.noda.tus.ac.jp

ABSTRACT

This paper proposes a block-based spatio-temporal prediction method for video coding. In this method, a predicted value at each pel is generated by a linear 3D predictor which uses the causal neighborhood in both the current and motion-compensated previous frames. When the causal neighborhood is within the block to be predicted, previously predicted values instead of the reconstructed ones are recursively used. Therefore, it can be incorporated with DCT-based residual coding algorithms where the reconstructed values are obtained on a block-by-block basis. In order to minimize the sum of squared prediction errors, a set of 3D predictors is iteratively optimized using the quasi-Newton method. Simulation results indicate that joint use of spatio-temporal prediction attains higher PSNR than exclusive use of spatial or temporal prediction in a framework of the proposed method.

1. INTRODUCTION

The latest video coding standard, H.264/AVC [1], employs various coding tools to achieve high coding performance. Some of the coding tools significantly contributing to the performance are connected with inter-frame prediction techniques. Those include variable block-size, quarter-pel accuracy and multiple reference frames for motion estimation and compensation. Another powerful tool is the intra-frame prediction technique which uses the reconstructed values of the adjacent blocks in the same frame. These facts indicate that exploitation of both spatial and temporal correlations through the prediction is quite useful for efficient video coding. However, the H.264/AVC alternatively switches the intra-frame and inter-frame prediction on a block-by-block basis, namely, joint use of spatial and temporal prediction is not allowed within a macroblock.

The effectiveness of joint spatio-temporal prediction has been already demonstrated in the area of lossless video coding. In [2], a linear 3D predictor which uses both the current and motion compensated reference frames is optimized pel-by-pel to enable efficient lossless video coding. In order to reduce computational complexity required at the decoder side, we have developed a block-adaptive prediction technique where a set of 3D predictors is optimized only at the encoder side and the resulting prediction coefficients are encoded as side information [3]. In both cases, actual prediction and succeeding entropy coding processes are performed in raster scan order. Therefore, the 3D predictor can always utilize the reconstructed values at spatially neighboring pels.

On the other hand, most of the current lossy video coding schemes employ block-based coding algorithm using DCT (Discrete Cosine Transform). This means the prediction process must also be a block-based one. Such a coding structure makes it difficult to exploit both spatial and temporal correlations simultaneously in the prediction process because there are no reconstructed values of spatially neighboring pels except at block boundaries. In fact, there are only a few studies attempting to use both spatial and temporal correlations at once in a framework of the block-based coding algorithm. In [4], unknown values inside of a block are fragmentarily copied from already reconstructed regions in the current and previous frames through a template matching mechanism. Another spatio-temporal prediction method based on refinement of motion compensated signals using spatially surrounding reconstructed pels was proposed in [5]. These studies stand on a kind of pattern analysis approach and the resulting predicted values are not optimal in a sense of MMSE (Minimum Mean Squared Error).

In this paper, a new spatio-temporal prediction method which can be incorporated with the DCT-based residual coding algorithm is proposed. The method is quite similar to the block-adaptive prediction technique [3], but already predicted values instead of the reconstructed ones are recursively used when reference pels of the 3D predictor are within the block to be predicted. Due to this recursive structure, design of the MMSE predictors results in a non-linear optimization problem. Therefore, we employ the quasi-Newton method, which is a popular gradient-based optimization algorithm, to solve the problem.

2. BLOCK-BASED SPATIO-TEMPORAL PREDICTION

In this paper, we consider that the DCT-based residual coding algorithm is performed in each block of 8×8 pels. Consequently, the prediction process must also be carried out in each block of the same size. To meet this constraint, predicted values in the target block, which is indicated by $\boldsymbol{B}_n$ in Figure 1, are generated by recursively applying a 3D predictor in raster scan order within the block. A predicted value at a pel $\boldsymbol{p}_0$ in the block $\boldsymbol{B}_n$ is calculated by the following equation:

$$\hat{s}_t(\boldsymbol{p}_0) = \sum_{k=1}^{K_1} a_k \cdot \tilde{s}_t(\boldsymbol{p}_k) + \sum_{k=1}^{K_2} b_k \cdot s'_{t-1}(\boldsymbol{q}_{k-1} + \boldsymbol{v}), \qquad (1)$$

$$\tilde{s}_t(\boldsymbol{p}_k) = \begin{cases} \hat{s}_t(\boldsymbol{p}_k) & (\boldsymbol{p}_k \in \boldsymbol{B}_n) \\ s'_t(\boldsymbol{p}_k) & (\text{otherwise}) \end{cases}, \qquad (2)$$

© EURASIP, 2010 ISSN 2076-1465

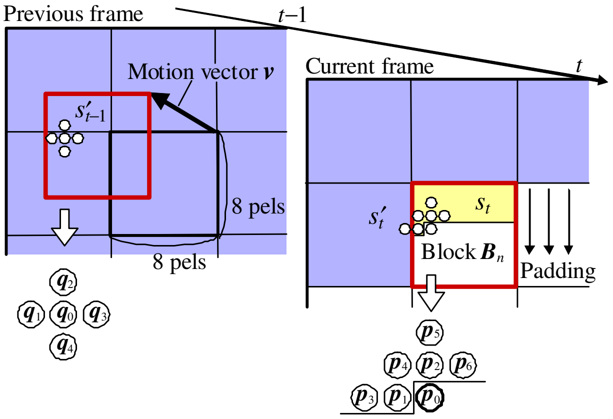


Figure 1: Block-based spatio-temporal prediction.

where $\{\boldsymbol{p}_k | k = 1, 2, \ldots, K_1\}$ and $\{\boldsymbol{q}_k | k = 0, 1, \ldots, K_2 - 1\}$ are reference pels disposed on the current and previous frames, respectively. Positions of the latter reference pels are motion-compensated according to a motion vector $\boldsymbol{v}$ which is detected for each macroblock of 16×16 pels. $\tilde{s}_t(\boldsymbol{p}_k)$ represents an image value used for spatial prediction. Since the reconstructed values $s'_t(\boldsymbol{p}_k)$ and $s'_{t-1}(\boldsymbol{q}_k)$ are available only in already encoded blocks, which are shown as colored blocks in Figure 1, a predicted value $\hat{s}_t(\boldsymbol{p}_k)$ in a causal neighborhood is used as $\tilde{s}_t(\boldsymbol{p}_k)$ if the reference pel $\boldsymbol{p}_k$ is inside of the block $\boldsymbol{B}_n$. In addition, when the reference pel $\boldsymbol{p}_k$ is in a right side block of $\boldsymbol{B}_n$, the corresponding reconstructed value is copied from the upper side (it is so called padding operation), and then used as $\tilde{s}_t(\boldsymbol{p}_k)$ for the prediction.

In this method, prediction coefficients $\{a_k, b_k\}$ play a crucial role. By appropriately changing their values, the 3D predictor provides various types of spatio-temporal prediction partially including the conventional intra-frame and inter-frame prediction. It should be noted that the motion vector ($\boldsymbol{v}$) is detected with integer-pel accuracy. However, our method has an ability to perform accurate motion-compensation because prediction coefficients $\{b_k\}$ give a similar effect to the adaptive interpolation filters [6] when order of temporal prediction (K_2) is sufficiently high. In this paper, multiple sets of prediction coefficients $\{a_k, b_k\}$ are optimized frame-by-frame, and the most suitable one is adaptively selected for each block. Both the optimization and the block-adaptive selection are carried out in an MMSE sense as described below.

3. OPTIMIZATION OF PREDICTION COEFFICIENTS

Let us consider that M kinds of 3D predictors, each of which has $K_1 + K_2$ prediction coefficients $\{a_1, \ldots, a_{K_1}, b_1, \ldots, b_{K_2}\}$, are assigned to the current frame on a block-by-block basis. Here we want to minimize the sum of squared prediction errors over the blocks where the same predictor is assigned. This is an unconstrained optimization problem with the following objective function:

$$J = \sum_{n \in \boldsymbol{\Omega}(m)} \sum_{\boldsymbol{p}_0 \in \boldsymbol{B}n} \left\{ s_t(\boldsymbol{p}_0) - \hat{s}_t(\boldsymbol{p}_0) \right\}^2, \tag{3}$$

where $s_t(\boldsymbol{p}_0)$ represents an original image value of the current frame and $\boldsymbol{\Omega}(m)$ is a set which consists of indices of the blocks sharing the m-th predictor. As mentioned above, the predicted value $\hat{s}_t(\boldsymbol{p}_0)$ is recursively calculated using the previously predicted values in the causal neighborhood. Therefore, the objective function J contains high order terms of the prediction coefficients and its minimization is formulated as a non-linear optimization problem with respect to variables $\{a_1, \ldots, a_{K_1}, b_1, \ldots, b_{K_2}\}$. To solve this problem, we employ the quasi-Newton method which is widely used for minimization of multi-variable non-linear functions [7]. The quasi-Newton method requires calculation of gradient vectors of the objective function. A component of the gradient vector is expressed as sum of products between prediction errors and partial differentials of the predicted values:

$$\frac{\partial J}{\partial a_i} = -2 \sum_{n \in \boldsymbol{\Omega}(m)} \sum_{\boldsymbol{p}_0 \in \boldsymbol{B}_n} \left\{ s_t(\boldsymbol{p}_0) - \hat{s}_t(\boldsymbol{p}_0) \right\} \cdot \frac{\partial \hat{s}_t(\boldsymbol{p}_0)}{\partial a_i}. \tag{4}$$

Furthermore, the partial differentials in the above equation can be converted to the following recurrence formulas:

$$\frac{\partial \hat{s}_t(\boldsymbol{p}_0)}{\partial a_i} = \tilde{s}_t(\boldsymbol{p}_i) + \sum_{k=1}^{K_1} a_k \cdot \frac{\partial \tilde{s}_t(\boldsymbol{p}_k)}{\partial a_i}, \tag{5}$$

$$\frac{\partial \tilde{s}_t(\boldsymbol{p}_k)}{\partial a_i} = \begin{cases} \partial \hat{s}_t(\boldsymbol{p}_k)/\partial a_i & (\boldsymbol{p}_k \in \boldsymbol{B}_n) \\ \partial s'_t(\boldsymbol{p}_k)/\partial a_i = 0 & (\text{otherwise}) \end{cases}. \tag{6}$$

In the same way, a gradient component with respect to the variables $\{b_i\}$ can be expressed as:

$$\frac{\partial J}{\partial b_i} = -2 \sum_{n \in \boldsymbol{\Omega}(m)} \sum_{\boldsymbol{p}_0 \in \boldsymbol{B}_n} \left\{ s_t(\boldsymbol{p}_0) - \hat{s}_t(\boldsymbol{p}_0) \right\} \cdot \frac{\partial \hat{s}_t(\boldsymbol{p}_0)}{\partial b_i}, \tag{7}$$

$$\frac{\partial \hat{s}_t(\boldsymbol{p}_0)}{\partial b_i} = s'_{t-1}(\boldsymbol{q}_{i-1} + \boldsymbol{v}) + \sum_{k=1}^{K_1} a_k \cdot \frac{\partial \tilde{s}_t(\boldsymbol{p}_k)}{\partial b_i}, \tag{8}$$

$$\frac{\partial \tilde{s}_t(\boldsymbol{p}_k)}{\partial b_i} = \begin{cases} \partial \hat{s}_t(\boldsymbol{p}_k)/\partial b_i & (\boldsymbol{p}_k \in \boldsymbol{B}_n) \\ \partial s'_t(\boldsymbol{p}_k)/\partial b_i = 0 & (\text{otherwise}) \end{cases}. \tag{9}$$

These equations show that not only the predicted values but also the gradient vectors required in the quasi-Newton method can be recursively calculated using the previously obtained results in the causal neighborhood. Strictly speaking, the reconstructed value $s'(\boldsymbol{p}_k)$ is not a constant when the reference pel $\boldsymbol{p}_k$ belongs to the block using the same predictor. Since the quasi-Newton method is performed in an iterative manner, use of the reconstructed values obtained at the previous iteration would be a practical solution.

4. DESIGN OF MULTIPLE PREDICTORS

In order to generate predicted values at the decoder side, the proposed method must encode the following parameters as side information.

- Motion vectors $\{\boldsymbol{v}\}$ with integer-pel accuracy detected in the respective macroblocks (16×16 pels).
- Predictor labels $\{m\}$ which specify the 3D predictors assigned to the respective blocks (8×8 pels).
- M sets of prediction coefficients $\{a_1, \ldots, a_{K_1}, b_1, \ldots, b_{K_2}\}$.

These parameters are iteratively optimized so that the sum of squared prediction errors calculated over the whole frame can be a minimum. Concrete procedures are as follows:

(1) Motion vectors $\{\boldsymbol{v}\}$ with integer-pel accuracy are detected by the block matching algorithm.

(2) Provisional predictor labels are assigned to all the blocks according to M level quantization of the sum of squared errors obtained by the above block matching.

(3) An initial predictor is designed for each region composed of blocks sharing the same predictor $\{\boldsymbol{B}_n \mid n \in \boldsymbol{\Omega}(m)\}$. In this step, MMSE predictors are designed on the assumption that the original image is available at causal neighbors. Therefore, the obtained 3D predictors are equivalent to the ones used in lossless video coding [3].

(4) The predictor label m is renewed for each block by selecting the optimal 3D predictor which minimizes the sum of squared prediction errors.

(5) The motion vector $\boldsymbol{v}$ is renewed for each macroblock within a search area of 3×3 pels.

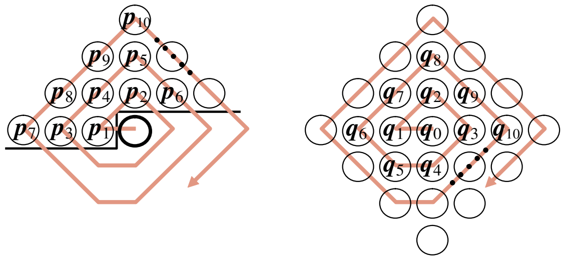
(6) Prediction coefficients $\{a_1, \ldots, a_{K_1}, b_1, \ldots, b_{K_2}\}$ of each 3D predictor are optimized using the quasi-Newton method.

(7) Procedures (4), (5) and (6) are iteratively carried out until all of the parameters converge.

The quasi-Newton method minimizes the objective function by iteratively performing the line search algorithm in a descent direction. The descent direction is determined based on an approximation of the Hessian matrix. In our implementation, the BFGS (Broyden-Fletcher-Goldfarb-Shanno) formula is employed for the Hessian matrix approximation using a series of previously calculated gradient vectors [7].

5. EXPERIMENTAL RESULTS

In order to evaluate basic performance of the proposed method, several experiments are conducted using CIF-sized monochrome video sequences (352×288 pels, 10 Hz, 15 frames). Since the DCT-based residual coding is currently not yet implemented, the JPEG encoded images with quality of about 35 dB are used in place of the reconstructed values $s'_t(\boldsymbol{p}_k)$ and $s'_{t-1}(\boldsymbol{q}_k)$ through the experiments. Reference pels $\{\boldsymbol{p}_k\}$ and $\{\boldsymbol{q}_k\}$ used for the spatio-temporal prediction are arranged in spiral order as shown in Figure 2. The parameters K_1 and K_2 are related to prediction order and their settings are crucial to obtain a good trade off between prediction performance and the amount of side information.



(a) Intra frame (b) Inter frame

Figure 2: Disposition of reference pels used for the proposed spatio-temporal prediction.

5.1 Spatial (intra-frame) prediction

The proposed method can be used for intra-coded pictures (I-pictures) by setting $K_2 = 0$. Figure 3 plots PSNRs of predicted images (averaged in 15 frames) as a function of prediction order K_1. In this figure, dotted lines indicate PSNRs achieved by the conventional intra-frame prediction method specified in High Profile of the H.264/AVC standard [8]. The method adaptively switches nine prediction modes and referred to as 'H.264/AVC 8×8 intra prediction'. It is shown that the proposed method outperforms the

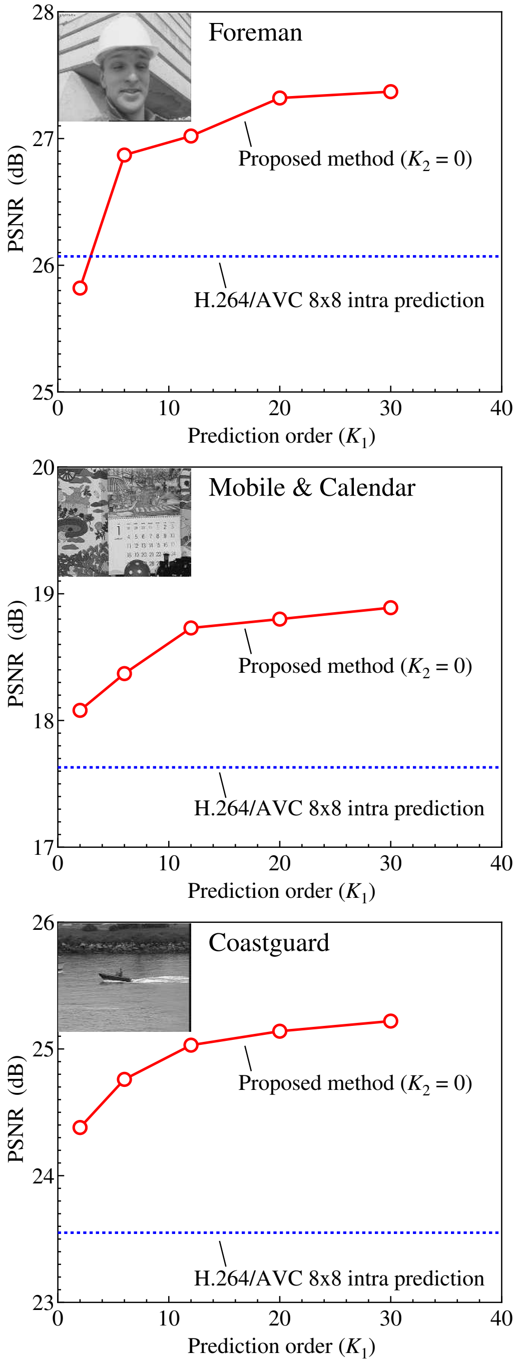


Figure 3: PSNRs of predicted images for I-pictures vs. prediction order K_1 ($M = 10$, $K_2 = 0$).

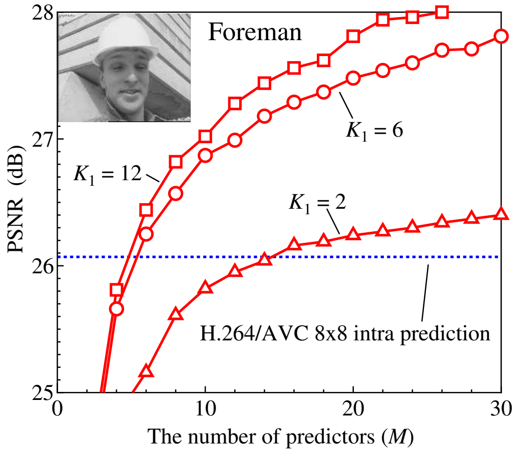
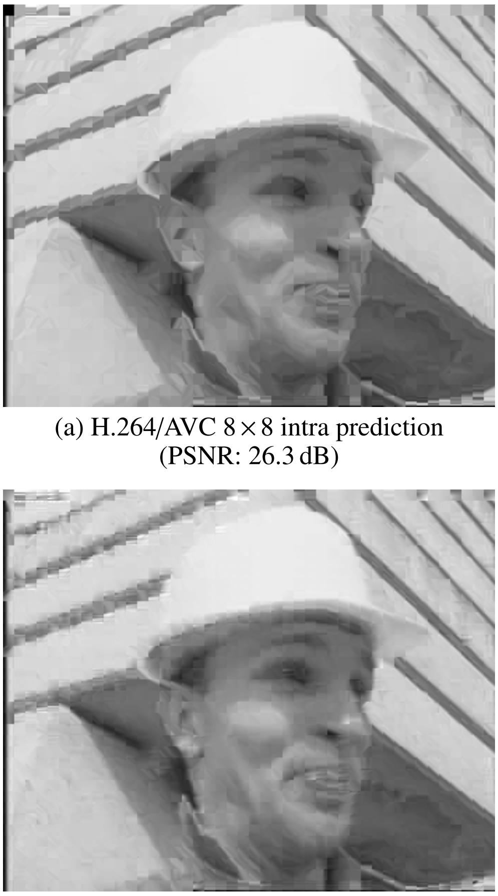


Figure 4: PSNRs of predicted images for I-pictures vs. the number of predictors M ($K_2 = 0$).



(a) H.264/AVC 8×8 intra prediction (PSNR: 26.3 dB)

(b) Proposed method ($M = 10$, $K_1 = 6$, $K_2 = 0$, PSNR: 27.0 dB)

Figure 5: Comparison of predicted images (I-picture).

conventional intra-frame prediction method except at $K_1 = 2$ for 'Foreman'. The effect of varying the number of 3D predictors (M) is shown in Figure 4. When enough prediction order is given ($K_1 \geq 6$), the proposed method provides higher PSNRs with fewer varieties of prediction ($M \geq 6$) than the H.264/AVC 8×8 intra prediction which has nine prediction modes. Moreover, the advantage of the proposed method in terms of visual quality is demonstrated in Figure 5.

5.2 Spatio-temporal prediction

Figure 6 plots PSNRs of predicted images for predictive-coded pictures (P-pictures) as a function of prediction order $K_1 + K_2$ under the condition of fixed number of 3D predictors ($M = 10$). In this figure, dotted lines show PSNRs obtained by block matching based motion compensated prediction where a motion vector with quarter-pel accuracy is searched for each macroblock (16×16 pels). To obtain predicted values at sub-pel positions, six-tap and bilinear interpolation filters are employed in the same way as the H.264/AVC [1]. Moreover, data points connected by dashed lines indicate the proposed method using temporal only prediction ($K_1 = 0$). In that case, minimization of the sum of squared prediction errors is easily accomplished by solving linear normal equations and the obtained predictors are equivalent to the adaptive interpolation filters used for the motion-compensated prediction method proposed in [6]. It is observed that joint spatio-temporal prediction attains higher PSNRs than the temporal only prediction with the exception of $K_2 = 1$ for 'Mobile & Calendar'. The horizontal axis of the Figure 6 roughly indicates the amount of side information needed for the prediction coefficients. From this point of view, combinations of $K_1 = 2, 6$ and $K_2 = 5$ seem to be reasonable for video coding application. Finally, predicted images obtained by the proposed method with the conditions of $K_1 = 0, K_2 = 13$ and $K_1 = 6, K_2 = 5$ are shown in Figure 7. We can see that visual quality of a facial area with complicated motions is considerably improved by joint use of spatio-temporal prediction.

6. CONCLUSIONS

In this paper, we have proposed a block-based spatio-temporal prediction method for video coding application. The method can exploit spatial and temporal correlations of video signals simultaneously and is suitable for DCT-based residual coding technique which are commonly used in the current video coding schemes. Moreover, it has a potential for integrating two prediction techniques: adaptive intra prediction and motion-compensated prediction using adaptive interpolation filters. Our study is still exploratory and evaluation of actual coding performance with quantization of the prediction coefficients should be conducted in the near future.

REFERENCES

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

[2] 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.

[3] I. Matsuda, T. Shiodera and S. Itoh, "Lossless Video Coding Using Variable Block-Size MC and 3D Prediction Optimized for Each Frame," in *Proc. of 12th European Signal Processing Conf. (EUSIPCO 2004)*, pp. 1967–1970, Vienna, Austria, Sep. 6–10, 2004.

[4] K. Sugimoto, M. Kobayashi, Y. Suzuki, S. Kato and C. S. Boon, "Inter Frame Coding with Template Matching

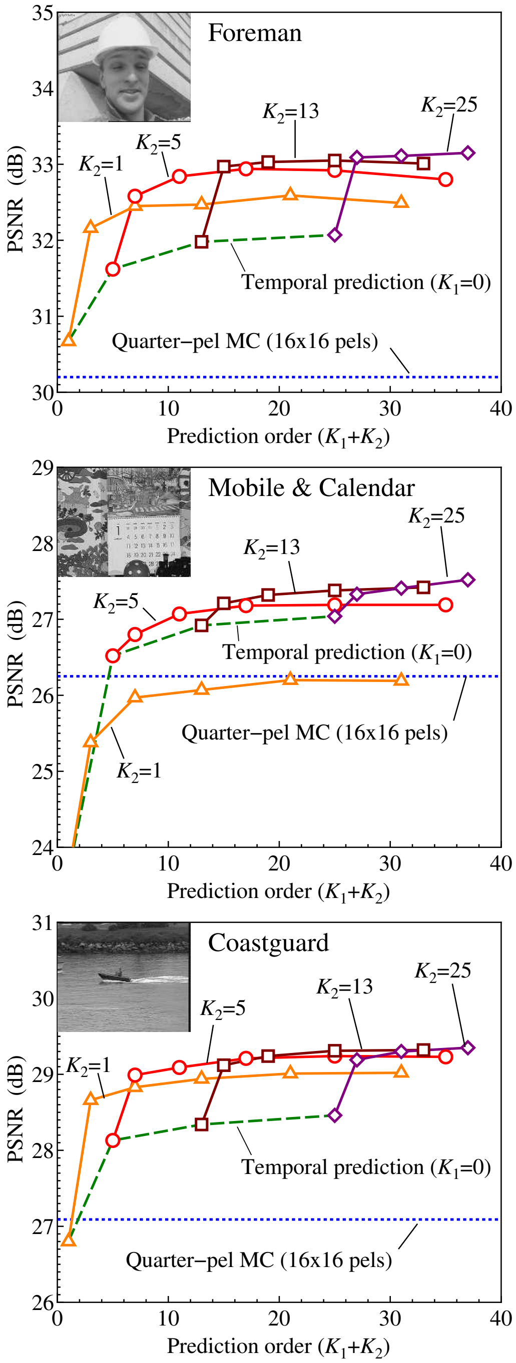
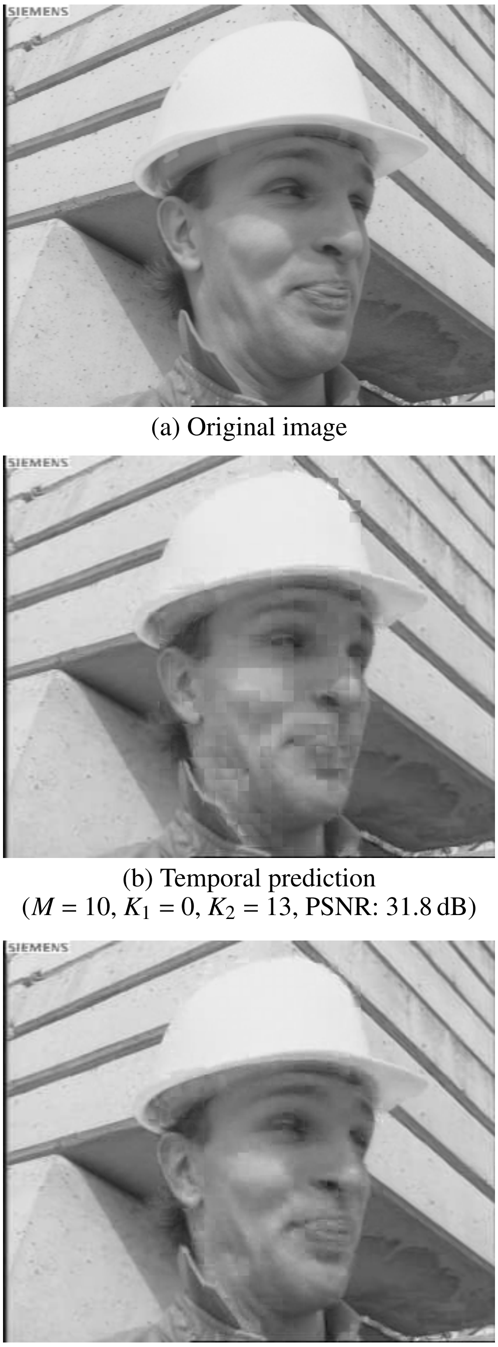


Figure 6: PSNRs of predicted images for P-pictures vs. prediction order $K_1 + K_2$ ($M = 10$).



(a) Original image

(b) Temporal prediction
(M = 10, K_1 = 0, K_2 = 13, PSNR: 31.8 dB)

(c) Spatio-temporal prediction
(M = 10, K_1 = 6, K_2 = 5, PSNR: 33.1 dB)

Figure 7: Examples of predicted images (P-picture).

Spatio-Temporal Prediction," *Proc. of 2004 IEEE International Conf. on Image Processing (ICIP 2004)*, pp. 465–468, Singapore, Oct. 24–27, 2004.

[5] J. Seiler and A. Kaup, "Spatio-Temporal Prediction in Video Coding by Spatially Refined Motion Compensation," in *Proc. of 2008 IEEE International Conf. on Image Processing (ICIP 2008)*, pp. 2788–2791, San Diego, California, Oct. 12–15, 2008.

[6] I. Matsuda, K. Yanagihara, S. Nagashima and S. Itoh, "Block Matching-Based Motion Compensation with Arbitrary Accuracy Using Adaptive Interpolation Filters," in *Proc. of 14th European Signal Processing Conf. (EUSIPCO 2006)*, Fri. 6.2, Florence, Italy, Sep. 4–8, 2006.

[7] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, "Numerical Recipes: The Art of Scientific Computing, Third Edition," Cambridge University Press, 2007.

[8] D. Marpe, T. Wiegand and S. Gordon, "H.264/MPEG4-AVC Fidelity Range Extensions: Tools, Profiles, Performance, and Application Areas," in *Proc. of 2005 IEEE International Conf. on Image Processing (ICIP 2005)*, Vol. 1, pp. 593–596, Geneva, Italy, Sep. 11–14, 2005.