Rate-Constrained Multihypothesis Prediction for Motion Compensated Video Compression

Markus Flierl, *Student Member, IEEE*, Thomas Wiegand, *Member, IEEE*, and
Bernd Girod, *Fellow, IEEE*

Abstract— **This article investigates linearly combined motion-compensated signals for video compression. In particular, we discuss multiple motion-compensated signals that are jointly estimated for efficient prediction and video coding. First, we extend the wide-sense stationary theory of motion-compensated prediction for the case of jointly estimated prediction signals. Our theory suggests that the gain by multihypothesis motion-compensated prediction is limited and that two jointly estimated hypotheses provide a major portion of this achievable gain. In addition, the analysis reveals a property of the displacement error of jointly estimated hypotheses. Second, we present a complete multihypothesis codec which is based on the ITU-T Recommendation H.263 with multiframe capability. Multiframe motion compensation chooses one prediction signal from a set of reference frames, whereas multihypothesis prediction chooses more than one for the linear combination. With our scheme, the time delay associated with B-frames is avoided by choosing more than one prediction signal from previously decoded pictures. Experimental results show that multihypothesis prediction improves significantly coding efficiency by utilizing variable block size and multiframe motion compensation. We show that variable block size and multihypothesis prediction provide gains for different scenarios and that multiframe motion compensation enhances the multihypothesis gain. For example, the presented multihypothesis codec with ten reference frames improves coding efficiency by up to 2.7 dB when compared to the reference codec with one reference frame for the set of investigated test sequences.**

Keywords— **Video Coding, Motion-Compensated Prediction, Rate-Constrained Motion Estimation, Multihypothesis Motion-Compensated Prediction, Linear Prediction, Multiframe Prediction, Entropy-Constrained Vector Quantization**

I. Introduction

TODAY'S state-of-the-art video codecs incorporate motion-compensated prediction (MCP). Some of these codecs employ more than one MCP signal simultaneously. The term "multihypothesis motion compensation" has been coined for this approach [1]. A linear combination of multiple prediction hypotheses is formed to arrive at the actual prediction signal. Theoretical investigations in [2] show that a linear combination of multiple prediction hypotheses can improve the performance of motion compensated prediction.

Bidirectional prediction for B-frames, as they are employed in H.263 [3] or MPEG [4], is an example for multihypothesis motion-compensated prediction where two motion-compensated signals are superimposed to reduce the bit-rate of a video codec. But the B-frame concept has to deal with a significant drawback: prediction uses the reference pictures before and after the B-picture. The associated delay may be unacceptable for interactive applications. To overcome this problem, the authors have previously proposed prediction algorithms [5], [6], [7] which superimpose multiple prediction signals from past frames only.

Selecting hypotheses from several past reference frames can be accomplished with the concept of long-term memory motion-compensated prediction [8] by extending each motion vector by a picture reference parameter. This concept is also called multiframe motion-compensated prediction [9]. The additional reference parameter overcomes the restriction that a specific hypothesis has to be chosen from a certain reference frame and enables the multihypothesis motion estimator to find an efficient set of prediction signals employing any of the reference frames. We will show that the concept of multiple reference frames also enhances the efficiency of multihypothesis video compression algorithms [10].

For bidirectional prediction, a joint estimation of forward and backward motion vectors is proposed in [11]. In that study, an iterative search procedure was used to estimates two motion vectors per block, but without a rate constraint. As the two motion vectors always point to the previous and subsequent frames, the advantage of the variable picture reference cannot be exploited. We generalize the joint estimation approach to several prediction signals, incorporate a rate constraint, and extend the motion vectors by picture reference parameters. Joint estimation is very efficient not only for bidirectional prediction but also for multihypothesis prediction with multiple reference frames [5].

Multihypothesis prediction allows the linear combination of an arbitrary number of prediction signals. In [2], the linear combination of motion-compensated prediction signals with statistically independent displacement errors is analyzed with the wide-sense stationary theory of motion-compensated prediction for hybrid video codecs. In this paper, we extend the wide-sense stationary theory to discuss the class of jointly estimated motion-compensated signals, their impact on displacement error correlation, and their performance bounds for arithmetic averaging. The joint estimation provides a set of complementary prediction signals with the property that their linear combination is very efficient for video compression. We show that combining two

M. Flierl is with the Telecommunications Laboratory, University of Erlangen-Nuremberg, Erlangen, Germany. He currently visits the Information Systems Laboratory at Stanford University. e-mail: flierl@LNT.de

T. Wiegand is with the Image Processing Department, Heinrich Hertz Institute, Berlin, Germany. e-mail: wiegand@hhi.de

B. Girod is with the Information Systems Laboratory, Stanford University, Stanford, CA, USA. e-mail: bgirod@stanford.edu

hypotheses already achieves most of the gain possible with multihypothesis motion-compensated prediction.

Multihypothesis motion-compensated prediction is not only a generalization of bidirectional prediction for B-frames. Overlapped block motion compensation (OBMC) [12], [13] fits also in this framework. OBMC is derived in [13] as a linear estimator of each pixel intensity, given that the only motion information available to the decoder is a set of block-based vectors. OBMC predicts the frame by overlapping shifted blocks of pixels from the reference frame, each weighted by an appropriate window. OBMC uses more than one motion vector for predicting the same pixel but does not increase the number of vectors per block. In contrast, our new scheme also uses more than one motion vector for the same pixel but also assigns more than one motion vector per block. We adopt the proposed design of the predictor coefficients for linear filtering, add a rate constraint, and relate the design to rate-constrained vector quantization [14], [15]. For OBMC, [1] and [13] propose also an iterative estimation search procedure for optimized motion estimation. Multihypothesis motion estimation can be regarded as a generalization of this algorithm.

Motion-compensated prediction with blocks of variable size improves the efficiency of video compression algorithms by adapting spatially displacement information [16], [17], [18]. Variable block size (VBS) prediction assigns more than one motion vector per macroblock but it uses just one motion vector for a particular pixel. We can improve this scheme and use more than one motion vector for the same pixel by utilizing multihypothesis motion-compensated prediction for blocks of any size. We will show that multihypothesis motion-compensated prediction with variable block size improves compression efficiency of VBS schemes [19].

ITU-T Recommendation H.263 utilizes a hybrid video coding concept with block-based motion-compensated prediction and DCT-based transform coding of the prediction error. P-frame coding of H.263 employs INTRA and INTER coding modes. Multihypothesis motion-compensated prediction for P-frame coding is enabled by new coding modes that are derived from H.263 INTER coding modes. Annex U of ITU-T Rec. H.263 allows multiframe motion-compensated prediction but does not provide multihypothesis capability. A combination of H.263 Annex U with B-frames leads to the concept of multihypothesis multiframe prediction. In this paper, we do not use H.263 B-frames as we discuss interpolative prediction for in-order encoding of sequences. H.263 B-frames can only be used for out-of-order encoding of sequences. Further, the presented concept of multihypothesis multiframe prediction is much more general than the B-frames in H.263. ITU-T Rec. H.263 also provides OBMC capability. As discussed previously, OBMC uses more than one motion vector for predicting the same pixel but those motion vectors are also used by neighboring blocks. In this work, a block predicted by multihypothesis motion-compensation has its individual set of motion vectors. We do not overlap shifted blocks that might be obtained by utilizing spatially neighboring motion vectors. The INTER4V coding mode of H.263 utilizes VBS prediction with either OBMC or an in-loop deblocking filter. An extension of H.263 OBMC or in-loop deblocking filter for multihypothesis prediction will go beyond the scope of this paper.

The outline of this article is as follows: In Section II, the concept of multihypothesis motion-compensated prediction is presented. Section III discusses a model for multihypothesis motion-compensated prediction and incorporates optimal multihypothesis motion estimation. This analysis provides insight about the number of hypotheses that have to be combined for an efficient video compression algorithm. Section IV integrates multihypothesis MCP into a H.263 video codec and discusses syntax extensions and coder control issues. Section V provides experimental results and demonstrates the efficiency of multihypothesis motion-compensated prediction for video coding. Further, the impact of variable block size and multiple reference prediction on the multihypothesis codec is investigated.

II. Multihypothesis Motion-Compensated Prediction

A. Multihypothesis Motion Compensation

Standard block-based motion compensation approximates each block in the current frame by a spatially displaced block chosen from the previous frame. As an extension, long-term memory motion compensation chooses the block from several previously decoded frames [8]. The motion-compensated signal is chosen by the transmitted motion vector and picture reference parameter.

Now, let us consider N motion-compensated signals. We will refer to them as hypotheses. The multihypothesis prediction signal is the linear superposition of these N hypotheses. Constant scalar coefficients determine the weight of each hypothesis for the predicted block. We will use only N scalar coefficients where each coefficient is applied to all pixel values of the corresponding hypothesis. That is, spatial filtering of hypotheses and OBMC are not employed.

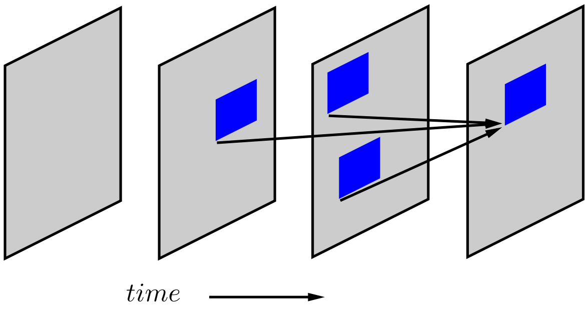

Fig. 1. Multihypothesis motion-compensated prediction with three hypotheses. Three blocks of previous decoded frames are linearly combined to form a prediction signal for the current frame.

Fig. 1 shows three hypotheses from previous decoded frames which are linearly combined to form the multihypothesis prediction signal for the current frame. Please note that a hypothesis can be chosen from any reference frame. Therefore, each hypothesis has to be assigned an individual picture reference parameter.

The proposed scheme differs from the concept of B-frame prediction in three significant ways: First, all reference frames are chosen from the past. No reference is made to a subsequent frame, as with B-frames, and hence no extra delay is incurred. Second, hypotheses are not restricted to stem from particular reference frames due to the picture reference parameter. This enables the encoder to find a much more accurate set of prediction signals, at the expense of a minor increase in the number of bits needed to select them. Third, it is possible to combine more than two motion-compensated signals. As will be shown later, these three properties of multihypothesis motion compensation improve the coding efficiency of a H.263 codec without incurring the delay that would be caused by using B pictures.

We strive to design the multihypothesis motion-compensated predictor in such a way that mean-squared prediction error is minimized while limiting the bit-rate consumed by the motion vectors and picture reference parameters. With variable length coding of the side information, the best choice of hypotheses will depend on the code tables used, while the best code tables depend on the probabilities of choosing certain motion vector/reference parameter combinations. Further, the best choice of hypotheses also depends on the linear coefficients used to weight each hypothesis, while the best coefficients depend on the covariance matrix of the hypotheses.

To solve this design problem, we find it useful to interpret multihypothesis motion-compensated prediction as a vector quantization problem. The *Generalized Lloyd Algorithm* [20] in conjunction with *Entropy Constrained Vector Quantization* [14], [21], [22] is employed to solve the design problem iteratively. For the interpretation, we argue that a block in the current frame is quantized. The output index of the quantizer is the index of the displacement vector. Each displacement vector is represented by a unique entropy codeword. Further, the codebook used for quantization contains motion-compensated blocks chosen from previous frames. This codebook is adaptive as the reference frames change with the current frame. For multihypothesis prediction, the codebook contains N-tuple of motion-compensated blocks whose components are linearly combined. This interpretation is sufficient to motivate a cost function for multihypothesis motion-compensated prediction.

Rate-constrained multihypothesis motion estimation utilizes a Lagrangian cost function. The costs are calculated by adding the mean-squared prediction error to a rate-term for the motion information, which is weighted by a Lagrange multiplier [23]. The estimator minimizes this cost function on a block basis to determine multiple displacement parameter. This corresponds to the biased nearest neighbor condition familiar from vector quantization with rate constraint. The multihypothesis decoder combines linearly more than one motion-compensated signal which are determined by multiple displacement parameter. The centroid condition determines the weighting coefficients if multiple motion-compensated signals are linearly combined. In [5], several video sequences are encoded to show that the centroid condition asks for averaging the multiple hypotheses. More details are given in [5].

B. Multihypothesis Motion Estimation

Multihypothesis motion compensation requires the estimation of multiple motion vectors and picture reference parameters. Best prediction performance is obtained when the N motion vectors and picture reference parameters are jointly estimated. This joint estimation would be computationally very demanding. Complexity can be reduced by an iterative algorithm which improves conditionally optimal solutions step by step [24], [5].

The *Hypothesis Selection Algorithm* (HSA) in [10] is such an iterative algorithm. The HSA minimizes the instantaneous Lagrangian costs for each block in the current frame and therefore performs rate-constrained multihypothesis motion estimation. The performance of the HSA depends on its initialization. The initial N-hypothesis is generated by repeating the optimal 1-hypothesis N times. This optimal 1-hypothesis is determined by rate-constrained motion estimation and minimizes the Lagrangian cost function. The initial N-hypothesis causes the same prediction error than the optimal 1-hypothesis but requires a higher bit-rate due to multiple displacements. Now, the iterative algorithms keeps $N-1$ hypotheses fixed and optimizes the remaining one by minimizing the multihypothesis cost function. The algorithm continues to determine these conditional optimal hypotheses until the multihypothesis cost function has converged. The iterative process employs conditional optimization to each of the N hypotheses. Further, all N hypotheses are compensated with the same motion accuracy and selected from the same search space. Consequently, there is no preference among the N hypotheses and all contribute equally.

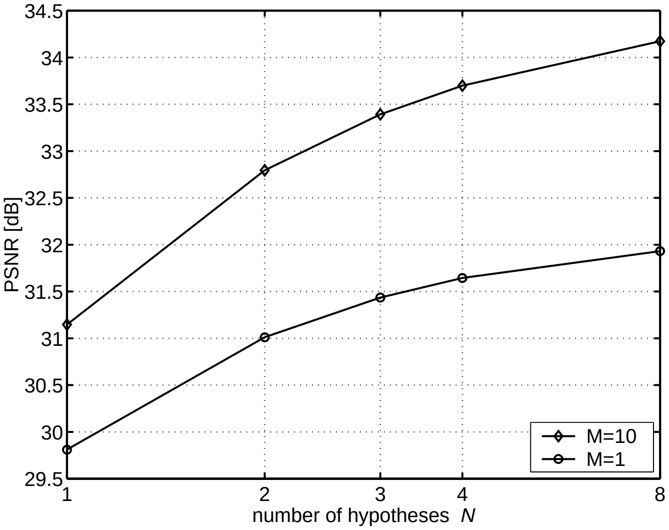

Fig. 2. Quality of the prediction signal and the number of hypotheses N for the sequence *Foreman* (QCIF, 10 fps, 10s), 16×16 blocks, half-pel accuracy, and no rate constraint. M indicates the number of reference frames.

Fig. 2 demonstrates the performance of HSA for equally weighted hypotheses and without a rate constraint, i.e., the Lagrange multiplier is set to zero. The quality of the prediction signal is plotted over the number of hypotheses N when predicting 16×16 blocks from past frames of the non-coded sequence *Foreman*. The quality of the prediction signal is given as average PSNR in dB. Half-pel accuracy is obtained by spatial bilinear interpolation. M is the number of frames that precede the current frame and are used for reference. It can be observed that increasing the number of hypotheses improves the quality of the prediction signal. Eight hypotheses on the previous reference frame ($M = 1$) improve the prediction signal approximately by 2 dB. The same number of hypotheses on ten previous reference frames ($M = 10$) achieve approximately 3 dB over single-hypothesis MCP with $M = 10$. Remarkably, multihypothesis MCP benefits from multiframe MCP such that the PSNR prediction gain is more than additive.

C. Rate-Distortion Performance

It is important to note that a N-hypothesis uses N motion vectors and picture reference parameters to form the prediction signal. Applying a product code for these N references will approximately increase the motion vector bit-rate for N-hypothesis MCP by factor of N. This higher rate has to be justified by the improved prediction quality.

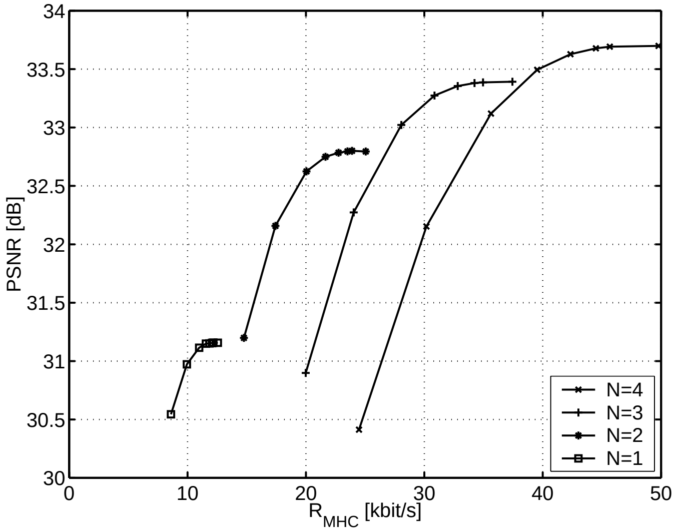

Fig. 3. Quality of the prediction signal vs. rate of the N-hypothesis code for the sequence *Foreman* (QCIF, 10 fps, 10s), 16×16 blocks, half-pel accuracy, and $M = 10$.

Fig. 3 depicts the quality of the prediction signal over the rate of the multihypothesis code R_{MHC} when predicting 16×16 blocks with $N = 1, 2, 3$, or 4 hypotheses from $M = 10$ past frames of the original sequence *Foreman*. The rate of the multihypothesis code is the number of bits used to code motion vectors and reference frame parameters. Half-pel accurate motion compensation is employed and utilizes spatial bilinear interpolation. The rate-PSNR points on each curve are obtained by varying the Lagrange multiplier. The quality without rate constraint (top right of each curve) is also depicted in Fig. 2 with $M = 10$. It can be observed that each predictor on its own is not the best one in the rate-distortion sense: For the same prediction quality, the one-hypothesis predictor provides always the lowest bit-rate. On the other hand, improved prediction quality can only be obtained for increasing number of hypotheses. It is shown in Section V that adaptively switching among the multihypothesis predictors improves the overall rate-distortion efficiency.

III. Efficient Number of Hypotheses

An efficient video compression algorithm should trade-off between complexity and achievable gain. The analysis in this section investigates this trade-off for multihypothesis prediction. We show that, first, the gain by multihypothesis MCP with averaged hypotheses is theoretically limited even if the number of hypotheses grows infinitely large and, second, two jointly estimated hypotheses provide a major portion of this achievable gain. The analysis is based on a power spectral model for inaccurate motion compensation [25], [26]. The work on multihypothesis prediction in [2] limits the discussion to statistically independent displacement errors between hypotheses. In the following, we extend this theory and allow statistically dependent displacement errors. We discuss the class of jointly estimated motion-compensated signals and their prediction performance bounds for arithmetic averaging. In particular, we focus on the dependency between multihypothesis prediction performance and displacement error correlation. A more detailed discussion of this theory is provided in [27].

A. Power Spectral Model for Inaccurate Multihypothesis Motion Compensation

Let $\mathbf{s}[l]$ and $\mathbf{c}_\mu[l]$ be scalar two-dimensional signals sampled on an orthogonal grid with horizontal and vertical spacing of 1. The vector $l = (x, y)^T$ denotes the location of the sample. For the problem of multihypothesis motion compensation, we interpret $\mathbf{c}_\mu$ as the μ-th of N motion-compensated signals available for prediction, and $\mathbf{s}$ as the current frame to be predicted. We call $\mathbf{c}_\mu$ also the μ-th hypothesis.

Obviously, multihypothesis motion-compensated prediction should work best if we compensate the true displacement of the scene exactly for each candidate prediction signal. Less accurate compensation will degrade the performance. To capture the limited accuracy of motion compensation, we associate a vector-valued displacement error $\mathbf{\Delta}_\mu$ with the μ-th hypothesis $\mathbf{c}_\mu$. The displacement error reflects the inaccuracy of the displacement vector used for motion compensation and transmission. The displacement vector field can never be completely accurate since it has to be transmitted as side information with a limited bit-rate. For simplicity, we assume that all hypotheses are shifted versions of the current frame signal $\mathbf{s}$. The shift is determined by the vector valued displacement error $\mathbf{\Delta}_\mu$ of the μ-th hypotheses. For that, the ideal reconstruction of the band-limited signal $\mathbf{s}[l]$ is shifted by the continuous valued displacement error and re-sampled on the original orthogonal grid. This translatory displacement model omits "noisy" signal components which are also included in [27].

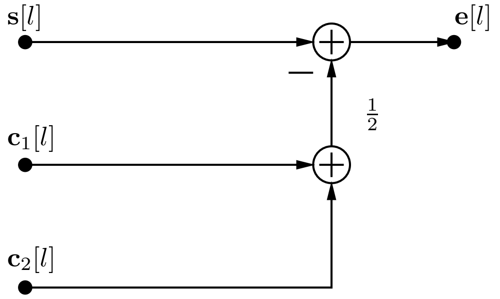

Fig. 4. Multihypothesis motion-compensated prediction with two hypotheses. The current frame $\mathbf{s}[l]$ is predicted by averaging two hypotheses $\mathbf{c}_1[l]$ and $\mathbf{c}_2[l]$.

Fig. 4 depicts the predictor which averages two hypotheses $\mathbf{c}_1[l]$ and $\mathbf{c}_2[l]$ in order to predict the current frame $\mathbf{s}[l]$. In general, the prediction error for each pel at location l is the difference between the current frame signal and N averaged hypotheses

$$\mathbf{e}[l] = \mathbf{s}[l] - \frac{1}{N}\sum_{\mu=1}^{N} \mathbf{c}_\mu[l]. \tag{1}$$

Assume that $\mathbf{s}$ and $\mathbf{c}_\mu$ are generated by a jointly wide-sense stationary random process with the real-valued scalar two-dimensional power spectral density $\Phi_{\mathbf{ss}}(\omega)$ as well as the cross spectral densities $\Phi_{\mathbf{c}_\mu\mathbf{s}}(\omega)$ and $\Phi_{\mathbf{c}_\mu\mathbf{c}_\nu}(\omega)$. Power spectra and cross spectra are defined according to

$$\Phi_{\mathbf{ab}}(\omega) = \mathcal{F}_*\left\{E\left\{\mathbf{a}[l_0+l]\mathbf{b}^*[l_0]\right\}\right\} \tag{2}$$

where $\mathbf{a}$ and $\mathbf{b}$ are complex signals, $\mathbf{b}^*$ is the complex conjugate of $\mathbf{b}$, and $l \in \Pi$ are the sampling locations. $\phi_{\mathbf{ab}}[l] = E\left\{\mathbf{a}[l_0+l]\mathbf{b}^*[l_0]\right\}$ is the scalar space-discrete cross correlation function between the signals $\mathbf{a}$ and $\mathbf{b}$ which (for wide-sense stationary random processes) does not depend on l_0 but only on the relative two-dimensional shift l. Finally, $\mathcal{F}_*\{\cdot\}$ is the 2D band-limited discrete-space Fourier transform

$$\mathcal{F}_*\left\{\phi_{\mathbf{ab}}[l]\right\} = \sum_{l\in\Pi} \phi_{\mathbf{ab}}[l] e^{-j\omega^T l} \quad \forall \quad \omega \in \,]-\pi,\pi]\times]-\pi,\pi] \tag{3}$$

where $\omega^T = (\omega_x, \omega_y)$ is the transpose of the vector valued frequency ω.

The power spectral density of the prediction error in (1) is determined by the power spectrum of the current frame and the cross spectra of the hypotheses

$$\Phi_{\mathbf{ee}}(\omega) = \tag{4}$$
$$\Phi_{\mathbf{ss}}(\omega) - \tfrac{2}{N}\sum_{\mu=1}^{N} \Re\left\{\Phi_{\mathbf{c}_\mu\mathbf{s}}(\omega)\right\} + \tfrac{1}{N^2}\sum_{\mu=1}^{N}\sum_{\nu=1}^{N} \Phi_{\mathbf{c}_\mu\mathbf{c}_\nu}(\omega),$$

where $\Re\{\cdot\}$ denotes the real component of the, in general, complex valued cross spectral densities $\Phi_{\mathbf{c}_\mu\mathbf{s}}(\omega)$. We adopt the expressions for the cross spectra from [2], where the displacement errors $\boldsymbol{\Delta}_\mu$ are interpreted as random variables which are statistically independent from $\mathbf{s}$:

$$\Phi_{\mathbf{c}_\mu\mathbf{s}}(\omega) = \Phi_{\mathbf{ss}}(\omega) E\left\{e^{-j\omega^T\boldsymbol{\Delta}_\mu}\right\} \tag{5}$$

$$\Phi_{\mathbf{c}_\mu\mathbf{c}_\nu}(\omega) = \Phi_{\mathbf{ss}}(\omega) E\left\{e^{-j\omega^T(\boldsymbol{\Delta}_\mu-\boldsymbol{\Delta}_\nu)}\right\} \tag{6}$$

Like in [2], we will assume a power spectrum $\Phi_{\mathbf{ss}}$ that corresponds to an exponentially decaying isotropic autocorrelation function with a correlation coefficient $\rho_{\mathbf{s}}$.

B. Model for the Probability Density Function of the Displacement Error

For $\boldsymbol{\Delta}_\mu$, a 2-D stationary normal distribution with variance $\sigma^2_{\boldsymbol{\Delta}}$ and zero mean is assumed where the x- and y-components are statistically independent. The displacement error variance is the same for all N hypotheses. This is reasonable because all hypotheses are compensated with the same accuracy. Further, the pairs $(\boldsymbol{\Delta}_\mu, \boldsymbol{\Delta}_\nu)$ are assumed to be jointly Gaussian random variables. The predictor design in [5] showed that there is no preference among the N hypotheses. Consequently, the correlation coefficient $\rho_{\boldsymbol{\Delta}}$ between two displacement error components $\boldsymbol{\Delta}_{x\mu}$ and $\boldsymbol{\Delta}_{x\nu}$ is the same for all pairs of hypotheses. We arrange the N individual displacement error components $\boldsymbol{\Delta}_{x\mu}$ with $\mu = 1, 2, \ldots, N$ to the vector of displacement errors for the x-component according to $\boldsymbol{\Delta}_x = (\boldsymbol{\Delta}_{x1}, \boldsymbol{\Delta}_{x2}, \ldots, \boldsymbol{\Delta}_{xN})^T$. With that, the above assumptions lead to the covariance matrix of the displacement error for the x-component

$$C_{\boldsymbol{\Delta}_x\boldsymbol{\Delta}_x} = \sigma^2_{\boldsymbol{\Delta}} \begin{pmatrix} 1 & \rho_{\boldsymbol{\Delta}} & \cdots & \rho_{\boldsymbol{\Delta}} \\ \rho_{\boldsymbol{\Delta}} & 1 & \cdots & \rho_{\boldsymbol{\Delta}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{\boldsymbol{\Delta}} & \rho_{\boldsymbol{\Delta}} & \cdots & 1 \end{pmatrix}. \tag{7}$$

The covariance matrix of the displacement error for the y-component is identical to that of the x-component. It is well known that the covariance matrix is nonnegative definite [28]. As a consequence, the correlation coefficient $\rho_{\boldsymbol{\Delta}}$ in (7) has the limited range

$$\frac{1}{1-N} \le \rho_{\boldsymbol{\Delta}} \le 1 \quad \text{for} \quad N = 2, 3, 4, \ldots, \tag{8}$$

which is dependent on the number of hypotheses N. To obtain this result, we solve $\det(C_{\boldsymbol{\Delta}_x\boldsymbol{\Delta}_x}) = 0$ as the covariance matrix is singular for the lower bound. In contrast to the work in [2], we do not assume that the displacement errors $\boldsymbol{\Delta}_\mu$ and $\boldsymbol{\Delta}_\nu$ are mutually independent for $\mu \neq \nu$.

These assumptions allow us to express the expected values in (5) and (6) in terms of the 2-D Fourier transform P of the continuous 2-D probability density function of the displacement error $\boldsymbol{\Delta}_\mu$.

$$\begin{aligned} E\left\{e^{-j\omega^T\boldsymbol{\Delta}_\mu}\right\} &= \int_{\mathcal{R}^2} p_{\boldsymbol{\Delta}_\mu}(\Delta) e^{-j\omega^T\Delta} d\Delta \\ &= e^{-\frac{1}{2}\omega^T\omega\sigma^2_{\boldsymbol{\Delta}}} \\ &= P(\omega, \sigma^2_{\boldsymbol{\Delta}}) \end{aligned} \tag{9}$$

The expected value in (6) contains differences of Gaussian random variables. It is well known that the difference of two Gaussian random variables is also Gaussian. As the two random variables have equal variance $\sigma^2_{\mathbf{\Delta}}$, the variance of the difference signal results as $\sigma^2 = 2\sigma^2_{\mathbf{\Delta}}(1-\rho_{\mathbf{\Delta}})$. Therefore, we obtain for the expected value in (6)

$$E\left\{e^{-j\omega^T(\mathbf{\Delta}_\mu - \mathbf{\Delta}_\nu)}\right\} = P\left(\omega, 2\sigma^2_{\mathbf{\Delta}}(1-\rho_{\mathbf{\Delta}})\right) \quad \text{for} \quad \mu \neq \nu. \tag{10}$$

For $\mu = \nu$, the expected value in (6) is equal to one. With that, we obtain for the power spectrum of the prediction error in (5):

$$\frac{\Phi_{\mathbf{ee}}(\omega)}{\Phi_{\mathbf{ss}}(\omega)} = \frac{N+1}{N} - 2P(\omega, \sigma^2_{\mathbf{\Delta}}) + \frac{N-1}{N} P\left(\omega, 2\sigma^2_{\mathbf{\Delta}}(1-\rho_{\mathbf{\Delta}})\right) \tag{11}$$

Setting $\rho_{\mathbf{\Delta}} = 0$ provides a result which is presented in [2].

C. Optimal Multihypothesis Motion Estimation

The displacement error correlation coefficient influences the performance of multihypothesis motion compensation. An optimal multihypothesis motion estimator will select sets of hypotheses that optimize the performance of multihypothesis motion compensation. In the following, we focus on the relationship between the prediction error variance

$$\sigma^2_{\mathbf{e}} = \frac{1}{4\pi^2} \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \Phi_{\mathbf{ee}}(\omega) d\omega \tag{12}$$

and the displacement error correlation coefficient. The prediction error variance is a useful measure because it is related to the minimum achievable transmission bit-rate [2].

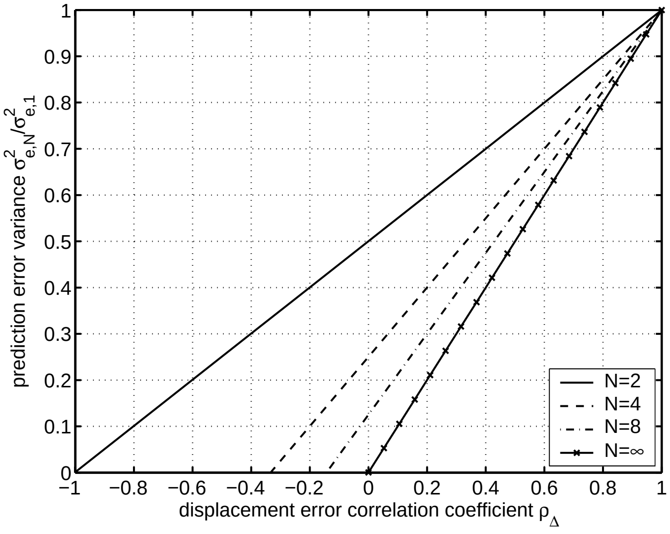

Fig. 5. Normalized prediction error variance for multihypothesis MCP over the displacement error correlation coefficient $\rho_{\mathbf{\Delta}}$. Reference is the single-hypothesis predictor. The hypotheses are averaged and no residual noise is assumed. The variance of the displacement error is set very small to $\sigma^2_{\mathbf{\Delta}} = 1/3072$.

Fig. 5 depicts the functional dependency of the normalized prediction error variance from the displacement error correlation coefficient $\rho_{\mathbf{\Delta}}$ within the range (8). The dependency is plotted for $N = 2, 4, 8$, and ∞ for very accurate motion compensation ($\sigma^2_{\mathbf{\Delta}} = 1/3072$). The correlation coefficient of the frame signal $\rho_{\mathbf{s}} = 0.93$ [2]. Reference is the prediction error variance of the single-hypothesis predictor $\sigma^2_{\mathbf{e},1}$. We observe that a decreasing correlation coefficient lowers the prediction error variance. (11) implies that this observation holds for any displacement error variance. Fig. 5 shows also that identical displacement errors ($\rho_{\mathbf{\Delta}} = 1$), and consequently, identical hypotheses will not reduce the prediction error variance compared to single-hypothesis motion compensation.

Without rate constraint, the optimal multihypothesis motion estimator minimizes not only the summed squared error but also its expected value [5]. If a stationary error signal is assumed, this optimal estimator minimizes the prediction error variance. $\sigma^2_{\mathbf{e}}$ increases monotonically for increasing $\rho_{\mathbf{\Delta}}$. This is a property of (11) which is also depicted in Fig. 5. The minimum of the prediction error variance is achieved for the lower bound of $\rho_{\mathbf{\Delta}}$. That is, an optimal multihypothesis motion estimator minimizes the prediction error variance by minimizing the displacement error correlation coefficient. Its minimum is given by the lower bound of the range (8).

$$\rho_{\mathbf{\Delta}} = \frac{1}{1-N} \quad \text{for} \quad N = 2, 3, 4, \ldots \tag{13}$$

This insight implies an interesting result for the case $N = 2$: Two jointly estimated hypotheses show the property that their displacement errors are maximally negatively correlated. The combination of two complementary hypotheses is more efficient than two hypotheses with independent displacement errors.

Let us consider the following one-dimensional example where the intensity signal is a continuous function of the spatial location x. A signal value that we want to use for prediction is given at spatial location $x = 0$. Due to an inaccurate displacement, only the signal value at spatial location $x = \Delta_1$ is available. We assume that the intensity signal is smooth around $x = 0$ and not spatially constant. When we pick the signal value at spatial location $x = \Delta_2 = -\Delta_1$ and average the two signal values we will get closer to the signal value at spatial location $x = 0$. Interpreting this as a random experiment, we get for the random variables $\mathbf{\Delta}_1 = -\mathbf{\Delta}_2$. This results in $\rho_{\mathbf{\Delta}} = -1$.

Fig. 6 depicts the rate difference for multihypothesis motion-compensated prediction over the displacement inaccuracy β for statistically independent displacement errors according to [2]. The rate difference [29], [2]

$$\Delta R = \frac{1}{8\pi^2} \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \log_2\left(\frac{\Phi_{\mathbf{ee}}(\omega)}{\Phi_{\mathbf{ss}}(\omega)}\right) d\omega \tag{14}$$

represents the maximum bit-rate reduction (in bit/sample) possible by optimum encoding of the prediction error $\mathbf{e}$, compared to optimum intra-frame encoding of the signal $\mathbf{s}$ for Gaussian wide-sense stationary signals for the same

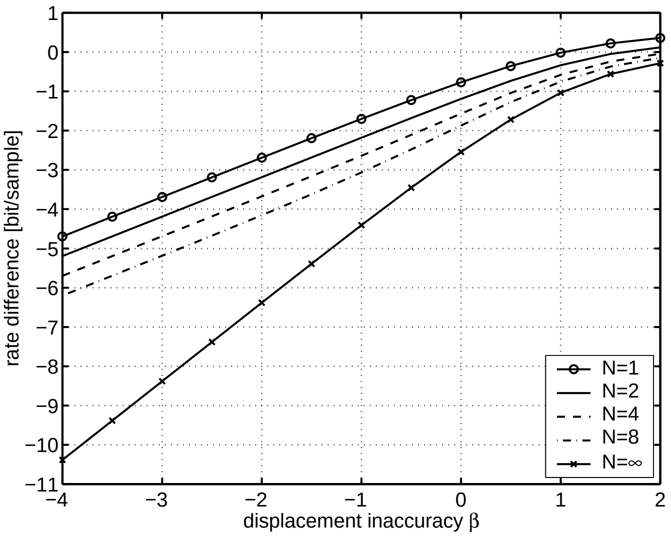

Fig. 6. Rate difference for multihypothesis MCP over the displacement inaccuracy β for statistically independent displacement errors. The hypotheses are averaged and no residual noise is assumed.

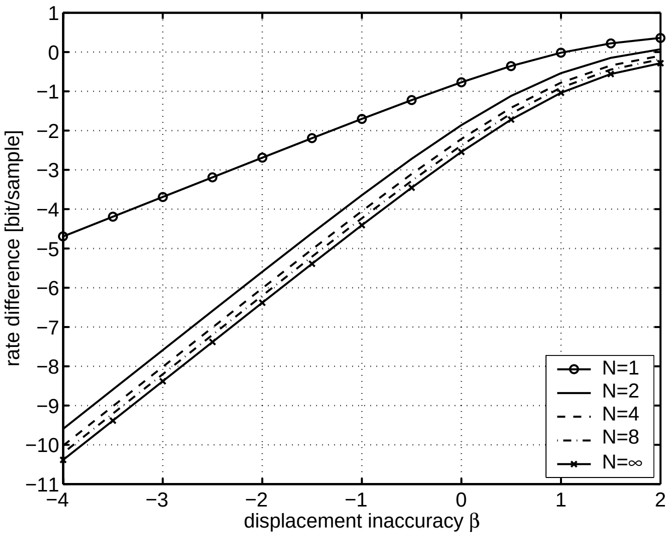

Fig. 7. Rate difference for multihypothesis MCP over the displacement inaccuracy β for optimized displacement error correlation. The hypotheses are averaged and no residual noise is assumed.

mean squared reconstruction error. A negative ΔR corresponds to a reduced bit-rate compared to optimum intraframe coding. The maximum bit-rate reduction can be fully realized at high bit-rates, while for low bit-rates the actual gain is smaller [2]. The horizontal axis in Fig. 6 is calibrated by $\beta = \log_2(\sqrt{12}\sigma_{\mathbf{\Delta}})$. It is assumed that the displacement error is entirely due to rounding and is uniformly distributed in the interval $[-2^{\beta-1}, 2^{\beta-1}] \times [-2^{\beta-1}, 2^{\beta-1}]$, where $\beta = 0$ for integer-pel accuracy, $\beta = -1$ for half-pel accuracy, $\beta = -2$ for quarter-pel accuracy, etc [2]. The displacement error variance is

$$\sigma^2_{\mathbf{\Delta}} = \frac{2^{2\beta}}{12}. \qquad (15)$$

We observe in Fig. 6 that doubling the number of hypotheses decreases the bit-rate up to 0.5 bit per sample and the slope reaches up to 1 bit per sample and inaccuracy step. The case $N \to \infty$ achieves a slope up to 2 bit per sample and inaccuracy step. This can also be observed in (11) for $N \to \infty$ when we apply a Taylor series expansion of second order for the function P.

$$\frac{\Phi_{\mathbf{ee}}(\omega)}{\Phi_{\mathbf{ss}}(\omega)} \approx \sigma^4_{\mathbf{\Delta}} \frac{1}{4} \left(\omega^T \omega\right)^2 \quad \text{for} \quad \sigma^2_{\mathbf{\Delta}} \to 0, N \to \infty, \rho_\Delta = 0 \qquad (16)$$

Inserting this result in (14) supports the observation in Fig. 6 for $N \to \infty$.

$$\Delta R \approx 2\beta + const. \quad \text{for} \quad \sigma^2_{\mathbf{\Delta}} \to 0, N \to \infty, \rho_\Delta = 0 \qquad (17)$$

Fig. 7 depicts the rate difference for multihypothesis motion-compensated prediction over the displacement inaccuracy β for optimized displacement error correlation according to (13). We observe for accurate motion compensation that the slope of the rate difference of 2 bit per sample and inaccuracy step is already reached for $N = 2$. For increasing number of hypotheses the rate difference converges to the case $N \to \infty$ at constant slope. This suggests that a practical video coding algorithm should utilize two jointly estimated hypotheses. Experimental results in Fig. 2 also suggest that the gain by multihypothesis prediction is limited and that two jointly estimated hypotheses provide a major portion of this achievable gain.

IV. Integration into H.263

The presented multihypothesis video codec is based on a standard hybrid video codec as proposed in ITU-T Recommendation H.263 [3]. Such a codec utilizes motion-compensated prediction to generate a prediction signal from previous reconstructed frames in order to reduce the bit-rate of the residual encoder. For block-based MCP, one motion vector and one picture reference parameter which address the reference block in a previous reconstructed frame are assigned to each block in the current frame.

The multihypothesis video codec additionally reduces the bit-rate of the residual encoder by improving the prediction signal. The improvement is achieved by combining linearly more than one motion-compensated prediction signal. For block-based multihypothesis MCP, more than one motion vector and picture reference parameter, which address a reference block in previous reconstructed frames, is assigned to each block in the current frame. These multiple reference blocks are linearly combined to form the block-based multihypothesis prediction signal.

The coding efficiency is improved at the expense of increased computational complexity for motion estimation at the encoder. But this disadvantage can be tackled by efficient estimation strategies like successive elimination [30]. At the decoder, a minor complexity increase is caused by the selection and combination of multiple prediction signals. Please note that not all macroblocks utilize multihypothesis MCP.

A. Syntax Extensions

The syntax of H.263 is extended such that multihypothesis motion compensation is possible. On the macroblock level, two new modes, INTER2H and INTER4H, are added which allow two or four hypotheses per macroblock, respectively. These modes are similar to the INTER mode of H.263. The INTER2H mode additionally includes an extra motion vector and frame reference parameter for the second hypothesis. The INTER4H mode incorporates three extra motion vectors and frame reference parameters. For variable block size prediction, the INTER4V mode of H.263 is extended by a multihypothesis block pattern. This pattern indicates for each 8×8 block the number of motion vectors and frame reference parameters. This mode is called INTER4VMH. The multihypothesis block pattern has the advantage that the number of hypotheses can be indicated individually for each 8×8 block. This allows the important case that just one 8×8 block can be coded with more than one motion vector and frame reference parameter. The INTER4VMH mode includes the INTER4V mode when the multihypothesis block pattern indicates just one hypothesis for all 8×8 blocks.

B. Coder Control

The coder control for the multihypothesis video codec utilizes rate-distortion optimization by Lagrangian methods. For that, the average Lagrangian costs of a macroblock, given the previous encoded macroblocks, are minimized.

$$J = D + \lambda R \tag{18}$$

The average costs J are constituted by the average distortion D and the weighted average bit-rate R. The weight, also called Lagrangian multiplier λ, is tied to the macroblock quantization parameter Q by the relationship [31]

$$\lambda = 0.85 Q^2. \tag{19}$$

This generic optimization method provides the encoding strategy for the multihypothesis encoder: Minimizing the instantaneous Lagrangian costs for each macroblock will minimize the average Lagrangian costs, given the previous encoded macroblocks.

H.263 allows several encoding modes for each macroblock. The one with the lowest Lagrangian costs will be selected for the encoding. This strategy is also called rate-constrained mode decision [32], [31].

The new multihypothesis modes include both multihypothesis prediction and prediction error encoding. The Lagrangian costs of the new multihypothesis modes have to be evaluated for rate-constrained mode decision. The distortion of the reconstructed macroblock is determined by the summed squared error. The macroblock bit-rate includes also the rate of all motion vectors and picture reference parameters. This allows the best trade-off between multihypothesis MCP rate and prediction error rate [33].

As already mentioned, multihypothesis MCP improves the prediction signal by spending more bits for the side-information associated with the motion-compensating predictor. But the encoding of the prediction error and its associated bit-rate also determines the quality of the reconstructed block. A joint optimization of multihypothesis motion estimation and prediction error encoding is far too demanding. But multihypothesis motion estimation independent of prediction error encoding is an efficient and practical solution. This solution is efficient if rate-constrained multihypothesis motion estimation, as explained before, is applied.

For example, the encoding strategies for the INTER and INTER2H modes are as follows: Testing the INTER mode, the encoder performs successively rate-constrained motion estimation for integer-pel positions and rate-constrained half-pel refinement. Rate-constrained motion estimation incorporates the prediction error of the video signal as well as the bit-rate for the motion vector and picture reference parameter. Testing the INTER2H mode, the encoder performs rate-constrained multihypothesis motion estimation. Rate-constrained multihypothesis motion estimation incorporates the multihypothesis prediction error of the video signal as well as the bit-rate for two motion vectors and picture reference parameters. Rate-constrained multihypothesis motion estimation is performed by the HSA which utilizes in each iteration step rate-constrained motion estimation to determine a conditional rate-constrained motion estimate. Given the obtained motion vectors and picture reference parameters for the INTER and INTER2H modes, the resulting prediction errors are encoded to evaluate the mode costs. The encoding strategy for the INTER4H mode is similar. For the INTER4VMH mode, the number of hypotheses for each 8×8 block is determined after encoding its residual error.

V. Experimental Results

The multihypothesis codec is based on the ITU-T Recommendation H.263 [3] with unrestricted motion vector mode, four motion vectors per macroblock, and enhanced reference picture selection in sliding window buffering mode. In contrast to H.263, a joint entropy code for horizontal and vertical motion vector data as well as an entropy code for the picture reference parameter is used. The efficiency of the reference codec is comparable to those of the H.263 test model TMN-10 [34]. The test sequences are coded at QCIF resolution and 10 fps. Each sequence has a length of ten seconds. For comparison purposes, the PSNR values of the luminance component are measured and plotted over the total bit-rate for quantizer values of 4, 5, 7, 10, 15, and 25. The data of the first intra-frame coded picture, which is identical in all cases, is excluded from the results.

A. Multiple Hypotheses for Constant Block Size

We will investigate the coding efficiency of multihypothesis (MH) prediction with two and four hypotheses for constant block size. Figs. 8 and 9 depict the average luminance PSNR from reconstructed frames over the overall bit-rate for the sequences *Foreman* and *Mobile & Calendar*. The performance of the codec with baseline prediction (BL), multihypothesis prediction with

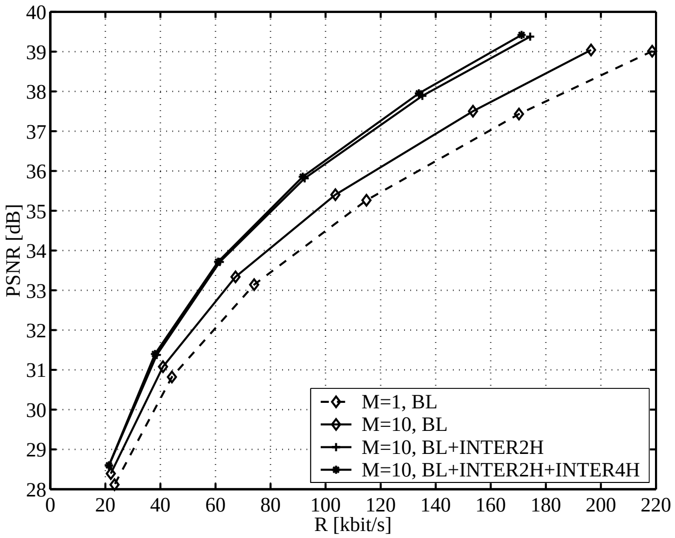

Fig. 8. Average luminance PSNR over total rate for the sequence *Foreman* depicting the performance of the multihypothesis coding scheme for constant block size. $M = 10$ reference pictures are utilized for prediction.

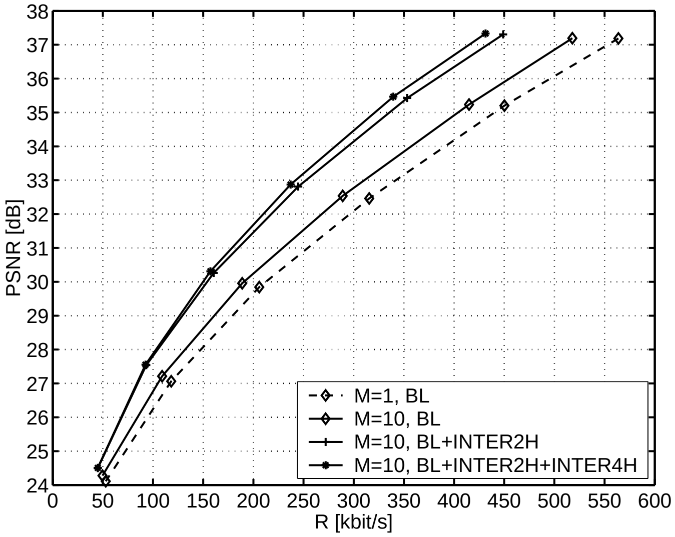

Fig. 9. Average luminance PSNR over total rate for the sequence *Mobile & Calendar* depicting the performance of the multihypothesis coding scheme for constant block size. $M = 10$ reference pictures are utilized for prediction.

two hypotheses (BL+INTER2H), and four hypotheses (BL+INTER2H+INTER4H) is shown. In each case, $M = 10$ reference pictures are utilized for prediction. The baseline performance for single frame prediction ($M = 1$) is added for reference.

Multihypothesis prediction is enabled by allowing the INTER2H mode on the macroblock level. A gain of up to 1 dB for the sequence *Foreman* and 1.4 dB for the sequence *Mobile & Calendar* is achieved by the INTER2H mode. Multihypothesis prediction with up to four hypotheses is adaptively implemented. A rate-distortion efficient codec should utilize four hypotheses only when their coding gain is justified by the associated bit-rate. In the case that four hypotheses are not efficient, the codec should be able to select two hypotheses and choose the INTER2H mode. The additional INTER4H mode gains just up to 0.1 dB for the sequence *Foreman* and 0.3 dB for the sequence *Mobile & Calendar*. This results also support the finding in Section III that two hypotheses provide the largest relative gain. Consequently, we will restrict our multihypothesis coding scheme to two hypotheses also considering the associated complexity for estimating four hypotheses.

B. Multiple Hypotheses for Variable Block Size

In this section, we investigate the influence of variable block size (VBS) prediction on multihypothesis prediction for $M = 10$ reference pictures. VBS prediction in H.263 is enabled by the INTER4V mode which utilizes four motion vectors per macroblock. VBS prediction is related to MH prediction in the way that more than one motion vector per macroblock is transmitted to the decoder. But both concepts provide gains for different scenarios. This can be verified by applying MH prediction to blocks of size 16×16 (INTER2H) as well as 8×8 (INTER4VMH). As we permit a maximum of two hypotheses per block, one bit is sufficient to signal whether one or two prediction signals are used.

Figs. 10 and 11 depict the average luminance PSNR from reconstructed frames over the overall bit-rate for the sequences *Foreman* and *Mobile & Calendar*. The performance of the codec with baseline prediction (BL), VBS prediction (BL+VBS), multihypothesis prediction with two hypotheses (BL+MHP(2)), and multihypothesis prediction with variable block size (BL+VBS+MHP(2)) is shown. In each case, $M = 10$ reference pictures are utilized for prediction. The baseline performance for single frame prediction ($M = 1$) is added for reference.

The combination of multihypothesis and variable block size prediction yields superior compression efficiency. For example, to achieve a reconstruction quality of 35 dB in PSNR, the sequence *Mobile & Calendar* is coded in baseline mode with 403 kbit/s for $M = 10$ (See Fig. 11). Correspondingly, MH prediction with $M = 10$ reduces the bitrate to 334 kbit/s. We save about 17% of the bit-rate for MH prediction on macroblocks. Performing MH prediction additionally on 8×8 blocks, the rate of the stream is 290 kbit/s in contrast to 358 kbit/s for the codec with VBS. MH prediction saves about 19% of the bit-rate produced by our codec with VBS prediction. Similar observations can be made for the sequence *Foreman* at 120 kbit/s. MH prediction on macroblocks gains about 1 dB over baseline prediction for $M = 10$ (See Fig. 10). Performing MH prediction additionally on 8×8 blocks, the gain is about 0.9 dB compared to the codec with VBS and $M = 10$ reference pictures.

Please note that the coding efficiency for the sequences *Foreman* (Fig. 10) and *Mobile & Calendar* (Fig. 11) is comparable for VBS prediction (BL+VBS) and MH prediction with two hypotheses (BL+MHP(2)) over the range of bitrates considered. MH prediction utilizes just two motion vectors and picture reference parameters compared to four for the INTER4V mode.

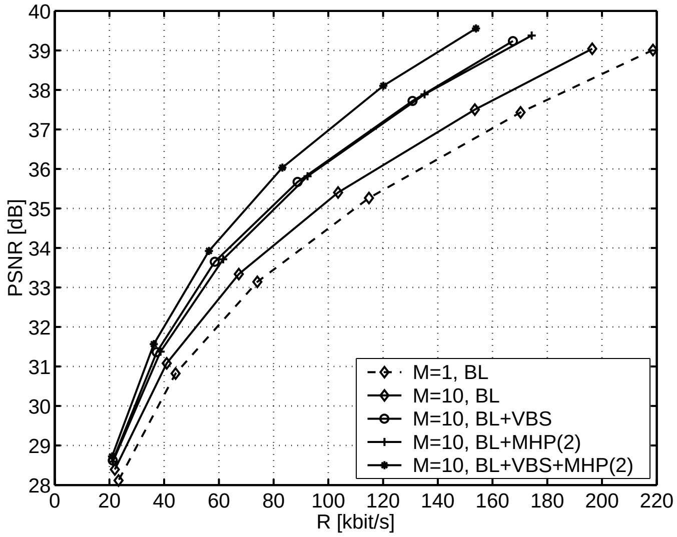

Fig. 10. Average luminance PSNR over total rate for the sequence *Foreman*. Multihypothesis and variable block size prediction can be successfully combined for compression. $M = 10$ reference pictures are utilized for prediction.

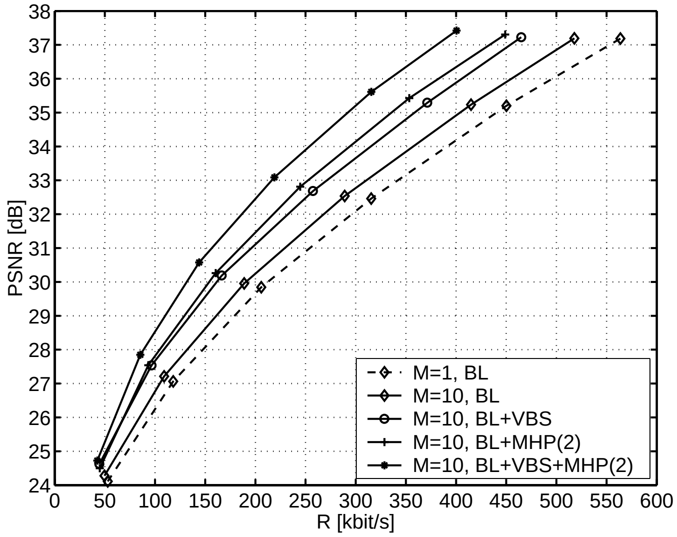

Fig. 11. Average luminance PSNR over total rate for the sequence *Mobile & Calendar*. Multihypothesis and variable block size prediction can be successfully combined for compression. $M = 10$ reference pictures are utilized for prediction.

For variable block size prediction, four hypotheses provide also no significant improvement over two hypotheses. For example, the multihypothesis codec with VBS and four hypotheses achieves just up to 0.3 dB gain over the codec with two hypotheses for the sequence *Mobile & Calendar*.

In summary, MH prediction works efficiently for both 16×16 and 8×8 blocks. The savings due to MH prediction are observed in the baseline mode as well as in the VBS prediction mode. Hence, our hypothesis selection algorithm is able to find two prediction signals on $M = 10$ reference frames which are combined more efficiently than just one prediction signal from these reference frames.

C. Multiple Hypotheses and Multiple Reference Pictures

The results presented so far are obtained for multihypothesis motion-compensated prediction with $M = 10$ reference pictures in sliding window buffering mode. In this section, the influence of long-term memory on the multihypothesis codec is investigated. It is demonstrated that two hypotheses chosen only from the prior decoded frame also improve coding efficiency. Additionally, the use of multiple reference frames enhances the efficiency of the multihypothesis codec.

Figs. 12 and 13 show the bit-rate savings at 35 dB of the decoded luminance signal over the number of reference frames M for the sequences *Foreman* and *Mobile & Calendar*. We compute PSNR vs. bit-rate curves by varying the quantization parameter and interpolate intermediate points by a cubic spline. The performance of the codec with variable block size prediction (VBS) is compared to the multihypothesis codec with two hypotheses (VBS+MHP(2)). Results are depicted for frame memory $M = 1, 2, 5, 10,$ and 20.

The multihypothesis codec with $M = 1$ reference frame has to choose both prediction signals from the previous decoded frame. The multihypothesis codec with VBS saves 7% for the sequence *Foreman* and 9% for the sequence *Mobile & Calendar* when compared to the VBS codec with one reference frame. For $M > 1$, more than one reference frame is allowed for each prediction signal. The reference frames for both hypotheses are selected by the rate-constrained multihypothesis motion estimation algorithm. The picture reference parameter allows also the special case that both hypotheses are chosen from the same reference frame. The rate constraint is responsible for the trade-off between prediction quality and bit-rate. Going from one reference frame to $M = 20$, the multihypothesis codec with VBS saves 25% for the sequence *Foreman* and 31% for the sequence *Mobile & Calendar* when compared to the VBS codec with one reference frame. For the same number of reference frames, the VBS codec saves about 15% for both sequences. The multihypothesis codec with VBS benefits when being combined with long-term memory prediction so that the savings are more than additive. The bit-rate savings saturate for 20 reference frames for both sequences.

Figs. 14 and 15 depict the average luminance PSNR over the total bit-rate for the sequences *Foreman* and *Mobile & Calendar*. The multihypothesis codec with variable block size (VBS+MHP(2)) is compared to the variable block size codec (VBS) for $M = 1$ and $M = 20$ reference frames. We can observe in these figures that multihypothesis prediction in combination with long-term memory motion compensation achieves coding gains up to 1.8 dB for *Foreman* and 2.8 dB for *Mobile & Calendar*. It is also observed that the use of multiple reference frames enhances the efficiency of multihypothesis prediction for video compression.

Finally, Figs. 12 and 13 suggest that a frame memory of $M = 10$ provides a good trade-off between encoder complexity and compression efficiency for our multihypothesis codec. Fig. 16 compares the multihypothesis codec with variable block size and frame memory $M = 10$ to the ref-

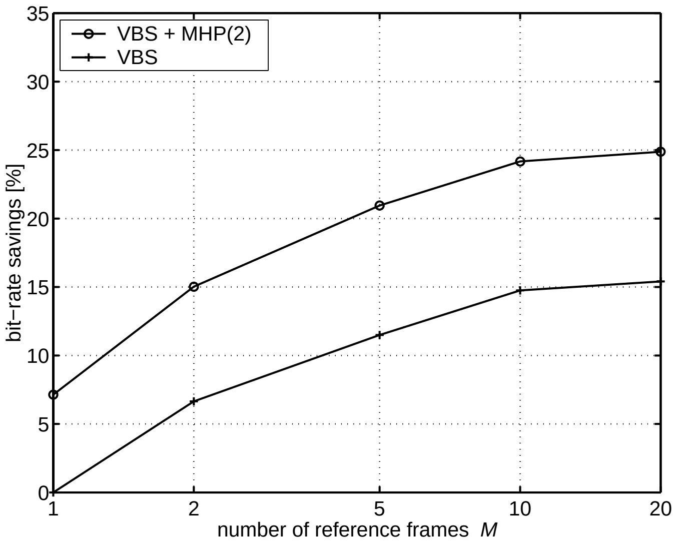

Fig. 12. Bit-rate savings at 35 dB PSNR over number of reference pictures for the sequence *Foreman*. The performance of the multihypothesis codec with variable block size is depicted for a variable number of reference frames M.

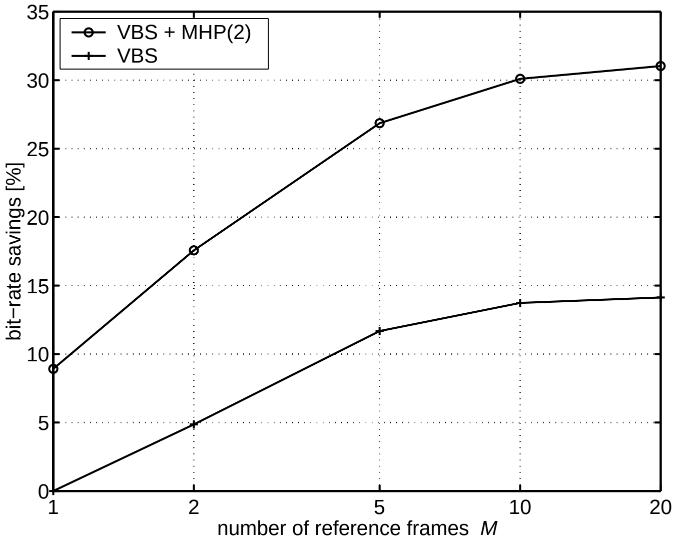

Fig. 13. Bit-rate savings at 35 dB PSNR over number of reference pictures for the sequence *Mobile & Calendar*. The performance of the multihypothesis codec with variable block size is depicted for a variable number of reference frames M.

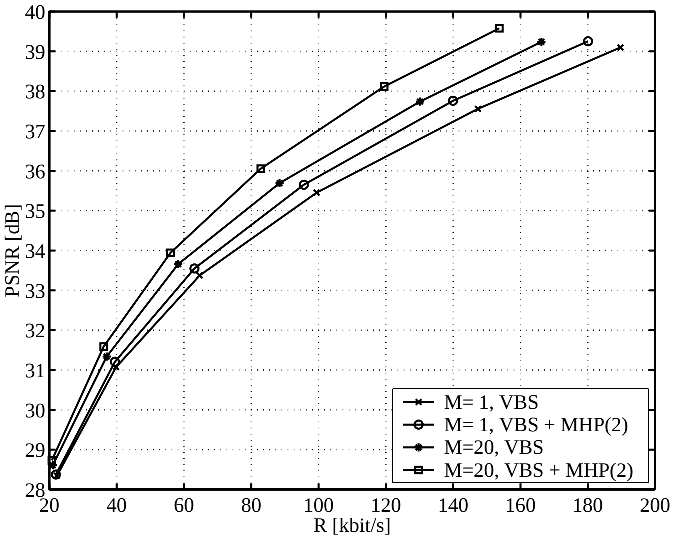

Fig. 14. Average luminance PSNR over total rate for the sequence *Foreman*. The performance of the multihypothesis codec with variable block size is depicted for $M = 1$ and $M = 20$ reference frames.

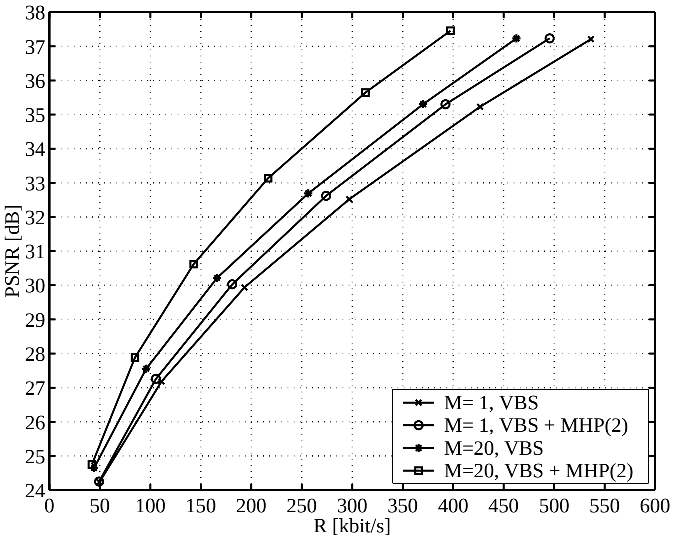

Fig. 15. Average luminance PSNR over total rate for the sequence *Mobile & Calendar*. The performance of the multihypothesis codec with variable block size is depicted for $M = 1$ and $M = 20$ reference frames.

erence codec with frame memory $M = 1$ and $M = 10$ for the sequences *Foreman* (top left), *Mobile & Calendar* (top right), *Sean* (bottom left), and *Weather* (bottom right). For each sequence the average luminance PSNR is depicted over the total bit-rate. The multihypothesis codec with long-term memory motion compensation achieves coding gains up to 1.8 dB for *Foreman*, 2.7 dB for *Mobile & Calendar*, 1.6 dB for *Sean*, and 1.5 dB for *Weather* compared to the reference codec with frame memory $M = 1$. The gain by long-term memory and multihypothesis prediction is comparable for the presented sequences.

VI. Conclusions

Motion-compensated prediction with multiple hypotheses improves the coding efficiency of state-of-the-art video compression algorithms by utilizing more than one motion vector and picture reference parameter per block to address multiple prediction signals. These signals are linearly combined with constant coefficients to form the prediction signal. Rate-constrained multihypothesis motion estimation is performed by the Hypothesis Selection Algorithm.

We have extended the wide-sense stationary theory of motion-compensated prediction with statistically independent hypotheses for the case of jointly estimated prediction signals and provided performance bounds for averaging

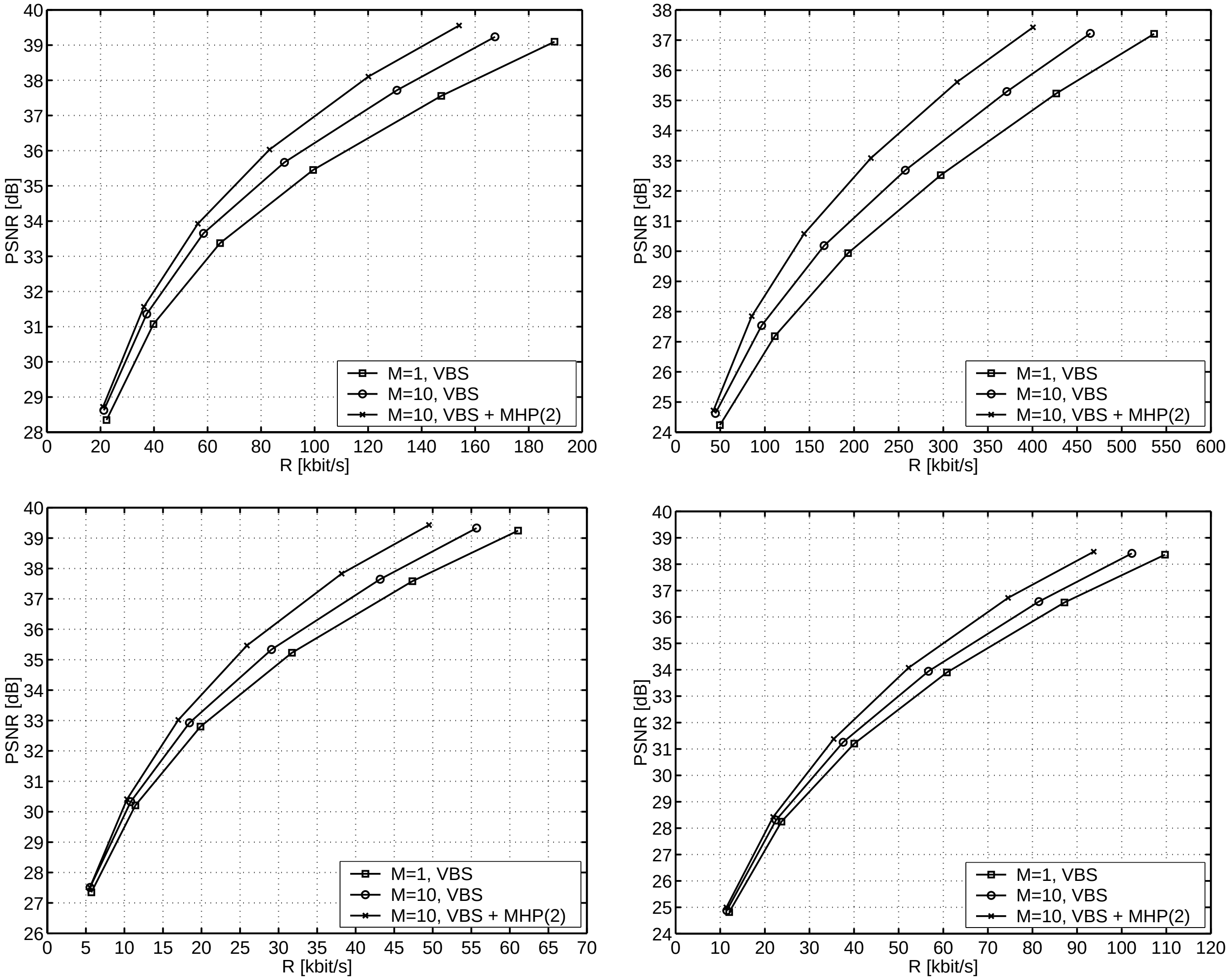

Fig. 16. Average luminance PSNR over total rate for the sequences *Foreman* (top left), *Mobile & Calendar* (top right), *Sean* (bottom left), and *Weather* (bottom right). The performance of the multihypothesis codec with variable block size and long-term memory motion compensation.

multiple hypotheses. The theory suggests that the gain by multihypothesis MCP with averaged hypotheses is limited even if the number of hypotheses is infinite. Two jointly estimated hypotheses provide a major portion of this achievable gain. The two hypotheses should possess respective displacement errors which are negatively correlated. Experimental results support this theoretical finding.

We present a complete multihypothesis codec which is based on the ITU-T Recommendation H.263 with variable block size and long-term memory motion compensation. In our experiments we observe that variable block size and multihypothesis prediction provide gains for different scenarios. Multihypothesis prediction works efficiently for both 16×16 and 8×8 blocks. Long-term memory enhances the efficiency of multihypothesis prediction. The multihypothesis gain and the long-term memory gain do not only add up; multihypothesis prediction benefits from hypotheses which can be chosen from different reference frames. Multihypothesis prediction with two hypotheses and long-term memory of ten frames achieves coding gains up to 2.7 dB, or equivalently, bit-rate savings up to 30% for the sequence *Mobile & Calendar* when compared to the reference codec with one frame memory. Therefore, multihypothesis prediction with long-term memory and variable block size is a very promising combination for efficient video compression.

References

[1] G.J. Sullivan, "Multi-hypothesis motion compensation for low bit-rate video coding", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Minneapolis, Apr. 1993, vol. 5, pp. 437–440.

[2] B. Girod, "Efficiency analysis of multihypothesis motion-compensated prediction for video coding", *IEEE Transactions on Image Processing*, vol. 9, no. 2, pp. 173–183, Feb. 2000.

[3] ITU-T, *Recommendation H.263, Version 2 (Video Coding for Low Bitrate Communication)*, 1998.

[4] A.M. Tekalp, *Digital Video Processing*, Prentice Hall, London, 1995.

[5] M. Flierl, T. Wiegand, and B. Girod, “A locally optimal design algorithm for block-based multi-hypothesis motion-compensated prediction”, in *Proceedings of the Data Compression Conference*, Snowbird, Utah, Apr. 1998, pp. 239–248.

[6] T. Wiegand, M. Flierl, and B. Girod, “Entropy-constrained linear vector prediction for motion-compensated video coding”, in *Proceedings of the International Symposium on Information Theory*, Cambridge, MA, Aug. 1998, p. 409.

[7] B. Girod, T. Wiegand, E. Steinbach, M. Flierl, and X. Zhang, “High-order motion compensation for low bit-rate video”, in *Proceedings of the European Signal Processing Conference*, Island of Rhodes, Greece, Sept. 1998.

[8] T. Wiegand, X. Zhang, and B. Girod, “Long-term memory motion-compensated prediction”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 70–84, Feb. 1999.

[9] M. Budagavi and J.D. Gibson, “Multiframe block motion compensated video coding for wireless channels”, in *Thirtieth Asilomar Conference on Signals, Systems and Computers*, Nov. 1996, vol. 2, pp. 953–957.

[10] M. Flierl, T. Wiegand, and B. Girod, “Rate-constrained multihypothesis motion-compensated prediction for video coding”, in *Proceedings of the IEEE International Conference on Image Processing*, Vancouver, Canada, Sept. 2000, vol. III, pp. 150–153.

[11] S.-W. Wu and A. Gersho, “Joint estimation of forward and backward motion vectors for interpolative prediction of video”, *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 684–687, Sept. 1994.

[12] S. Nogaki and M. Ohta, “An overlapped block motion compensation for high quality motion picture coding”, in *Proceedings of the IEEE International Symposium on Circuits and Systems*, May 1992, pp. 184–187.

[13] M.T. Orchard and G.J. Sullivan, “Overlapped block motion compensation: An estimation-theoretic approach”, *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 693–699, Sept. 1994.

[14] P.A. Chou, T. Lookabaugh, and R.M. Gray, “Entropy-constrained vector quantization”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 31–42, Jan. 1989.

[15] A. Gersho, “Optimal nonlinear interpolative vector quantization”, *IEEE Transactions on Communications*, vol. 38, no. 9, pp. 1285–1287, Sept. 1990.

[16] G.J. Sullivan and R.L. Baker, “Rate-distortion optimized motion compensation for video compression using fixed or variable size blocks”, in *Proceedings of the IEEE Global Telecommunications Conference*, Phoenix, AZ, Dec. 1991, vol. 3, pp. 85–90.

[17] J. Ribas-Corbera and D. L. Neuhoff, “On the optimal block size for block-based, motion-compensated video coders”, in *Proceedings of the SPIE Conference on Visual Communications and Image Processing*, San Jose, CA, Jan. 1997, vol. 2, pp. 1132–1143.

[18] T. Wiegand, M. Flierl, and B. Girod, “Entropy-constrained design of quadtree video coding schemes”, in *Proceedings of the International Conference on Image Processing and its Applications*, Dublin, Ireland, July 1997, vol. 1, pp. 36–40.

[19] M. Flierl, T. Wiegand, and B. Girod, “A video codec incorporating block-based multi-hypothesis motion-compensated prediction”, in *Proceedings of the SPIE Conference on Visual Communications and Image Processing*, Perth, Australia, June 2000, vol. 4067, pp. 238–249.

[20] M.J. Sabin and R.M. Gray, “Global convergence and empirical consistency of the generalized lloyd algorithm”, *IEEE Transactions on Information Theory*, vol. 32, no. 2, pp. 148–155, Mar. 1986.

[21] Y. Shoham and A. Gersho, “Efficient bit allocation for an arbitrary set of quantizers”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 9, pp. 1445–1453, Sept. 1988.

[22] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Press, 1992.

[23] H. Everett III, “Generalized lagrange multiplier method for solving problems of optimum allocation of resources”, *Operations Research*, vol. 11, pp. 399–417, 1963.

[24] J. Besag, “On the statistical analysis of dirty pictures”, *J. Roy. Statist. Soc. B*, vol. 48, no. 3, pp. 259–302, 1986.

[25] B. Girod, “The efficiency of motion-compensating prediction for hybrid coding of video sequences”, *IEEE Journal on Selected Areas in Communications*, vol. SAC-5, no. 7, pp. 1140–1154, Aug. 1987.

[26] B. Girod, “Motion-compensating prediction with fractional-pel accuracy”, *IEEE Transactions on Communications*, vol. 41, no. 4, pp. 604–612, Apr. 1993.

[27] M. Flierl and B. Girod, “Multihypothesis motion estimation for video coding”, in *Proceedings of the Data Compression Conference*, Snowbird, Utah, Mar. 2001, pp. 341–350.

[28] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1991.

[29] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[30] W. Li and E. Salari, “Successive elimination algorithm for motion estimation”, *IEEE Transactions on Image Processing*, vol. 4, no. 1, pp. 105–107, Jan. 1995.

[31] G.J. Sullivan and T. Wiegand, “Rate-distortion optimization for video compression”, *IEEE Signal Processing Magazine*, vol. 15, pp. 74–90, Nov. 1998.

[32] T. Wiegand, M. Lightstone, D. Mukherjee, T. G. Campbell, and S. K. Mitra, “Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 2, pp. 182–190, Apr. 1996.

[33] B. Girod, “Rate-constrained motion estimation”, in *Proceedings of the SPIE Conference on Visual Communications and Image Processing*, Chicago, Sept. 1994, pp. 1026–1034.

[34] ITU-T Video Coding Experts Group, *Video Codec Test Model, Near Term, Version 10 (TMN-10), Draft 1, Q15-D65*, Apr. 1998, http:// standards.pictel.com/ ftp/ video-site/ 9804_Tam/ q15d65.doc.

Markus Flierl (S'01) received the Dipl.-Ing. degree in electrical engineering from the University of Erlangen-Nuremberg, Germany, in 1997. From 1999 to 2001, he was a scholar with the Graduate Research Center at the University of Erlangen-Nuremberg. He is currently a visiting researcher with the Information Systems Laboratory, Stanford University, Stanford, CA. He contributed to the ITU-T Video Coding Experts Group standardization efforts. His current research interests are data compression, signal processing, and motion in image sequences.

Thomas Wiegand (S'92–M'93) received the Dr.-Ing. degree from University of Erlangen-Nuremberg, Germany, in 2000 and the Dipl.-Ing. degree in electrical engineering from Technical University of Hamburg-Harburg, Germany, in 1995. From 1993 to 1994, he was a Visiting Researcher at Kobe University, Japan. In 1995, he was a Visiting Scholar at the University of California at Santa Barbara, USA, where he started his research on video compression and transmission. Since then he has published several conference and journal papers on the subject and has contributed successfully to the ITU-T Video Coding Experts Group (ITU-T SG16 Q.6) standardization efforts. From 1997 to 1998, he has been a Visiting Researcher at Stanford University, USA, and served as a consultant to 8x8 (now Netergy Networks), Inc., Santa Clara, CA, USA. In cooperation with Netergy Networks, he holds two US patents in the area of video compression. In October 2000, he has been appointed as Associated Rapporteur of the ITU-T Video Coding Experts Group. In December 2001, he has been appointed as Associated Rapporteur / Co-Chair of the Joint Video Team (JVT) that has been created by ITU-T Video Coding Experts Group and ISO Moving Pictures Experts Group (ISO/IEC JTC1/SC29/WG11) for finalization of the H.26L project. His research interests include image compression, communication and signal processing as well as vision and computer graphics.

Bernd Girod (M'80–SM'97–F'98) is Professor of Electrical Engineering in the Information Systems Laboratory of Stanford University, California. He also holds a courtesy appointment with the Stanford Department of Computer Science. His research interests include networked multimedia systems, video signal compression, and 3-d image analysis and synthesis. He received his M.S. degree in Electrical Engineering from Georgia Institute of Technology, in 1980 and his Doctoral degree "with highest honours" from University of Hannover, Germany, in 1987. Until 1987 he was a member of the research staff at the Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung, University of Hannover, working on moving image coding, human visual perception, and information theory. In 1988, he joined Massachusetts Institute of Technology, Cambridge, MA, USA, first as a Visiting Scientist with the Research Laboratory of Electronics, then as an Assistant Professor of Media Technology at the Media Laboratory. From 1990 to 1993, he was Professor of Computer Graphics and Technical Director of the Academy of Media Arts in Cologne, Germany, jointly appointed with the Computer Science Section of Cologne University. He was a Visiting Adjunct Professor with the Digital Signal Processing Group at Georgia Institute of Technology, Atlanta, GA, USA, in 1993. From 1993 until 1999, he was Chaired Professor of Electrical Engineering/Telecommunications at University of Erlangen-Nuremberg, Germany, and the Head of the Telecommunications Institute I, co-directing the Telecommunications Laboratory. He has served as the Chairman of the Electrical Engineering Department from 1995 to 1997, and as Director of the Center of Excellence "3-D Image Analysis and Synthesis" from 1995-1999. He has been a Visiting Professor with the Information Systems Laboratory of Stanford University, Stanford, CA, during the 1997/98 academic year. As an entrepreneur, he has worked successfully with several start-up ventures as founder, investor, director, or advisor. Most notably, he has been a founder and Chief Scientist of Vivo Software, Inc., Waltham, MA (1993-98); after Vivo's acquisition, since 1998, Chief Scientist of RealNetworks, Inc. (Nasdaq: RNWK); and, since 1996, an outside Director of 8x8, Inc. (Nasdaq: EGHT). He has authored or co-authored one major text-book and over 200 book chapters, journal articles and conference papers in his field, and he holds about 20 international patents. He has served as on the Editorial Boards or as Associate Editor for several journals in his field, and is currently Area Editor for the IEEE Transactions on Communications as well as member of the Editorial Boards of the journals EURASIP Signal Processing, the IEEE Signal Processing Magazine, and the ACM Mobile Computing and Communication Review. He has chaired the 1990 SPIE conference on "Sensing and Reconstruction of Three-Dimensional Objects and Scenes" in Santa Clara, California, and the German Multimedia Conferences in Munich in 1993 and 1994, and has served as Tutorial Chair of ICASSP-97 in Munich and as General Chair of the 1998 IEEE Image and Multidimensional Signal Processing Workshop in Alpbach, Austria. He has been the Tutorial Chair of ICIP-2000 in Vancouver and the General Chair of the Visual Communication and Image Processing Conference (VCIP) in San Jose, CA, in 2001. He has been a member of the IEEE Image and Multidimensional Signal Processing Committee from 1989 to 1997 and was elected Fellow of the IEEE in 1998 'for his contributions to the theory and practice of video communications.' He has been named 'Distinguished Lecturer' for the year 2002 by the IEEE Signal Processing Society.