Motion-Compensating Prediction with Fractional-Pel Accuracy

Bernd Girod, *Member, IEEE*

Abstract— **The effect of fractional-pel accuracy on the efficiency of motion-compensating predictors is studied using various spatial prediction/interpolation filters. In model calculations, the power spectral density of the prediction error is related to the probability density function of the displacement error. Prediction can be improved both by higher accuracy of motion-compensation and by spatial Wiener filtering in the predictor. Beyond a critical accuracy, the possibility of further improving prediction by more accurate motion-compensation is small. Experiments with videophone signals and with broadcast TV signals confirm these model calculations. Sinc-interpolation, bilinear interpolation, and Wiener filtering are compared at integer-pel, 1/2-pel, 1/4-pel, and 1/8-pel accuracies. A new three-stage technique for reliable displacement estimation with fractional-pel accuracy is described; it is based on phase correlation. For motion-compensation with block size of 16 × 16 pels, quarter-pel accuracy appears to be sufficient for broadcast TV signals; while for videophone signals, half-pel accuracy is desirable.**

I. Introduction

For efficient transmission of time-varying images, motion-compensation is an important concept. It achieves data compression by exploiting the similarities between successive frames of a video signal. Motion-compensating prediction (MCP) is frequently proposed in the context of codecs for the transmission of limited-motion videotelephone or videoconferencing signals at medium and low bit rates, as well as for codecs accommodating full-motion material at higher rates [1]–[22]. The majority of motion-compensating predictors that are reported in the literature use motion-compensation with "integer-pel accuracy;" the contents of the previous picture are displaced by integer multiples of the horizontal and vertical sampling intervals. Since the true frame-to-frame displacements of the image contents are, of course, completely unrelated to the sampling grid, we expect improved prediction when motion-compensation with "fractional-pel accuracy" is employed. We will refer to this improvement as the "accuracy effect." Codecs that use motion-compensation with fractional-pel accuracy have been reported, e.g., in [6], [18], [19], and [21]. Typically, fractional-pel accuracy is achieved by simple bilinear interpolation, which produces a spatially blurred prediction signal. It has been pointed out that the use of a spatial lowpass filter in the predictor can improve prediction in connection with integer-pel accuracy [1], [6], [23]. Improvement gained in this way will be referred to as the "filtering effect." Bilinear interpolation for fractional-pel accuracy introduces spatial lowpass filtering as a side effect, and it has not been clear yet, how large contributions are from either the accuracy effect or the filtering effect.

Recently, a theoretical framework has been presented for analysis of the performance of MCP in terms of rate distortion theory [23]. It relates power spectral density of the prediction error to the accuracy of motion-compensation. It is shown in [23] that with integer-pel accuracy of the displacement estimate, the additional gain by MCP over optimum intraframe encoding of the signal is limited to ~0.8 bit/sample in moving areas. Larger gains require fractional-pel accuracy.

In this paper, we apply the theory developed in [23] to the analysis and the design of MCP, and complement it by experimental results. In particular, we will address the following questions.

- With fractional-pel accuracy MCP using a bilinear filter, how large are the contributions to improved prediction of the "accuracy effect" and of the "filtering effect"?
- To what extent can we further improve MCP by spatial filtering in the predictor?
- What kernel for interpolation to fractional-pel accuracy leads to the most efficient prediction?
- What accuracy of displacement measurement and motion-compensation is adequate?

These problems are investigated both theoretically and experimentally on the basis of video signals typical for videotelephone and broadcasting applications. In Section II, we briefly review hybrid coding of video signals using MCP and introduce some of the nomenclature used throughout this paper. Section III presents a model calculation which shows that beyond a certain "critical accuracy" the possibility of further improving prediction by more accurate motion-compensation is very small. In Section IV, a method for displacement estimation with fractional-pel accuracy is proposed. Finally, Section V experimentally compares different spatial prediction filters at integer-pel, 1/2-pel, 1/4-pel, and 1/8-pel accuracy of MCP.

II. Hybrid Coding of Video Signals Using Motion-Compensating Prediction

The structure of many proposed state-of-the-art codecs resembles the MCP hybrid coding scheme shown in Fig. 1 [1]–[23]. This coding scheme combines differential pulse code

Paper approved by The Editor for Image Processing of the IEEE Communications Society. Manuscript received September 18, 1989; revised April 12, 1990. This work was supported by an Ernst von Siemens fellowship. This paper was presented in part at the International Workshop on 64 kbit/s Coding of Moving Video, Hannover, Germany, 1988.

The author is with the Academy of Media Arts, Peter-Welter-Platz 2, 5000 Cologne 1 Germany.

IEEE Log Number 9209474.

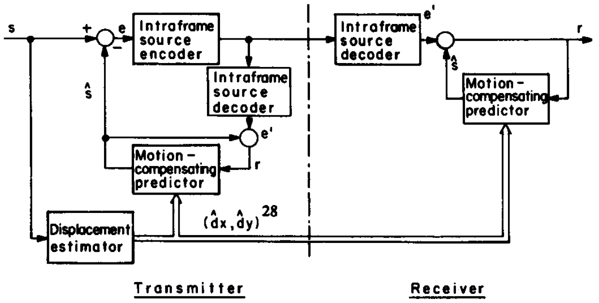

s
+
−
e
Intraframe source encoder
Intraframe source decoder
e'
ŝ
r
Motion-compensating predictor
Displacement estimator
$(\hat{d}x,\hat{d}y)$
28
Intraframe source decoder
e'
ŝ
Motion-compensating predictor
r
Transmitter
Receiver

Fig. 1. Block diagram of an MCP hybrid coding scheme.

modulation (DPCM) along an estimated motion trajectory of the picture contents with intraframe encoding of the prediction error e. The displacement estimate $(\hat{d}_x, \hat{d}_y)$ is transmitted in addition to the intraframe-encoded prediction error. At the receiver, the intraframe source decoder generates the reconstructed prediction error e', which differs from e by some source coding error. The transmitter contains a replication of the receiver in order to be able to generate the same prediction value $\hat{s}$.

In a simple scheme, the intraframe source encoder could be just a quantizer, such that the MCP hybrid coder (Fig. 1) is a motion-compensating DPCM coder [2], [3], [10], [14], [16]. More sophisticated schemes encode a spatial neighborhood of prediction errors e employing a blockwise transform, like the discrete cosine transform (DCT) [1], [5], [6], [8], [11]–[13], [15], [18], [19], a quadrature mirror filter bank [9], [17], vector quantization [4], [7], [20], or quadtree coding [22]. In each case, the intraframe encoder serves to remove spatial redundancy from the signal e and to adaptively reduce the spatial resolution, if the channel cannot accommodate a full-resolution signal. It has been pointed out by several authors that the motion-compensated prediction error signal is only weakly correlated spatially [11], [21]–[23]. Thus, the potential for redundancy reduction in the intraframe source encoder is very small. This finding suggests that the prediction error variance

$$\sigma_e^2 = E\{e^2\} - E^2\{e\} \tag{1}$$

is a useful measure that is directly related to the minimum achievable transmission bit rate for a given signal-to-noise ratio [23]. In (1), $E\{\cdot\}$ is the expectation operator. Throughout this paper, we will use σ_e^2 to evaluate the performance of MCP.

Fig. 2 shows the most general form of MCP. Let (x, y) be spatial coordinates. The prediction signal $\hat{s}(x, y)$ is obtained from the samples of the reconstructed, previous frame $r(x, y)$. MCP with fractional-pel accuracy is not straighforward, since $r(x, y)$ is only available at

$$(x_s, y_s) \in \Pi, \tag{2}$$

where Π is the set of sampling positions. Throughout this paper, we assume an orthogonal sampling grid with horizontal and vertical sampling intervals X and Y. For motion-compensation, $r(x, y)$ can be thought of as being interpolated to a space-continuous signal

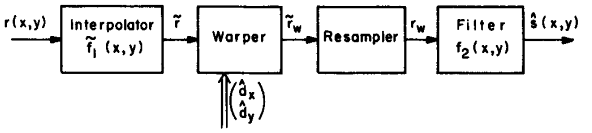


Fig. 2. Motion-compensating predictor.

$$\tilde{r}(x, y) = \sum_{(x_s, y_s)\in\Pi} \tilde{f}_I(x - x_s, y - y_s) \cdot r(x_s, y_s) \quad \forall\ (x, y) \in R^2, \tag{3}$$

where $\tilde{f}_I(x, y)$ is the interpolation filter kernel. Throughout this paper, space-continuous signals will be distinguished by the "tilde" from sampled, space-discrete signals. As immediate exception from this notational rule, let us introduce an estimated displacement $(\hat{d}_x(x, y), \hat{d}_y(x, y))$ defined over continuous spatial coordinates $(x, y) \in R^2$. The space-continuous $\tilde{r}(x, y)$ can be warped according to $(\hat{d}_x, \hat{d}_y)$ to yield the space-continuous signal

$$\tilde{r}_w(x, y) = \tilde{r}\Big(x - \hat{d}_x(x, y), y - \hat{d}_y(x, y)\Big) \quad \forall\ (x, y) \in R^2, \tag{4}$$

which is then resampled to

$$r_w(x, y) = \tilde{r}_w(x, y) \quad \forall\ (x, y) \in \Pi. \tag{5}$$

Of course, in a practical implementation, we would not calculate $\tilde{r}(x, y)$ and $\tilde{r}_w(x, y)$ continuously, but only at the positions that are required to obtain the resampled signal $r_w(x, y)$. Also, because of (5), $(\hat{d}_x(x, y), \hat{d}_y(x, y))$ is only required at $(x, y) \in \Pi$. In order to improve prediction, we finally convolve $r_w(x, y)$ with a space-discrete impulse response $f_2(x, y)$ and obtain the prediction

$$\hat{s}(x, y) = \sum_{(x_s, y_s)\in\Pi} r_w(x_s, y_s) \cdot f_2(x - x_s, y - y_s) \quad \forall\ (x, y) \in \Pi. \tag{6}$$

III. Theoretical Analysis of Motion-Compensating Prediction

The motion-compensating predictor (Fig. 2) is a linear, space-varying system. A general analysis is difficult. However, we can treat MCP with the traditional theory of linear, space-invariant systems if the estimated displacement $(\hat{d}_x, \hat{d}_y)$ does not depend on (x, y). Many MCP schemes employ a constant estimated displacement vector within a block of, e.g., 16 × 16 samples, and our assumption would hold within a block. Other schemes interpolate a sparse estimated vector field to obtain a dense, smooth field for motion-compensation. With the assumption of constant $(\hat{d}_x, \hat{d}_y)$, we neglect effects caused by spatial changes in the estimated vector field.

Combining (3), (4), and (5) yields

$$r_w(x, y) = \sum_{(x_s, y_s)\in\Pi} \tilde{f}_I\Big(x - x_s - \hat{d}_x, y - y_s - \hat{d}_y\Big) \cdot r(x_s, y_s) \quad \forall\ (x, y) \in \Pi. \tag{7}$$

If $(\hat{d}_x, \hat{d}_y)$ is constant, then (7) is a convolution of $r(x_s, y_s)$ with a sampled version of the interpolation kernel $\tilde{f}_1(x,y)$, with the sampling phase depending on $(\hat{d}_x, \hat{d}_y)$. Further treatment is conveniently carried out in the frequency domain by applying a *2-D band-limited Fourier transform*

$$H(\omega_x, \omega_y) = \sum_{(x,y)\in\Pi_s} h(x,y) e^{-j\omega_x \frac{x}{X} - j\omega_y \frac{y}{Y}} \qquad \forall\ |\omega_x| < \pi, \qquad |\omega_y| < \pi\,, \tag{8}$$

where Π_s is an arbitrary set of sampling positions forming an orthogonal grid with horizontal sampling interval X and vertical sampling interval Y. We do not require the origin $(x = 0, y = 0)$ to coincide with one of the samples. Thus, (8) is slightly more general than the conventional definition of the Fourier transform (e.g., see [24]). We restrict the region of support of the Fourier transform to the baseband, and do not consider baseband replications. This restriction greatly simplifies the following mathematics without sacrificing generality.

With definition (8), we can write (7) as

$$R_w(\omega_x, \omega_y) = e^{-j\omega_x \frac{\hat{d}_x}{X} - j\omega_y \frac{\hat{d}_y}{Y}} \cdot F_1(\omega_x, \omega_y) \cdot R(\omega_x, \omega_y)\,, \tag{9}$$

where $R_w(\omega_x, \omega_y)$ and $R(\omega_x, \omega_y)$ are band-limited Fourier transforms (8) of $r_w(x,y)$ and $r(x,y)$, respectively. $F_1(\omega_x,\omega_y)$ is the Fourier transform (8) of the sampled interpolation kernel

$$f_1(x,y) = \tilde{f}_1(x,y) \quad \forall\ \left(x - \hat{d}_x, y - \hat{d}_y\right) \in \Pi\,. \tag{10}$$

The condition $(x - \hat{d}_x, y - \hat{d}_y) \in \Pi$ indicates that $f_1(x,y)$ is defined only at positions (x,y), where $(x - \hat{d}_x, y - \hat{d}_y)$ coincides with a sampling position in the grid Π (2), which is the sampling grid of the reconstructed signal $r(x,y)$. With a definition of the combined spatial filtering characteristic of the predictor,

$$F(\omega_x, \omega_y) = F_1(\omega_x, \omega_y) \cdot F_2(\omega_x, \omega_y)\,, \tag{11}$$

we can fully describe the motion-compensating predictor by

$$\hat{S}(\omega_x, \omega_y) = e^{-j\omega_x \frac{\hat{d}_x}{X} - j\omega_y \frac{\hat{d}_y}{Y}} \cdot F(\omega_x, \omega_y) \cdot R(\omega_x, \omega_y)\,, \tag{12}$$

where $\hat{S}(\omega_x, \omega_y)$ is the Fourier transform (8) of $\hat{s}(x,y)$.

We now assume that the input video signal $s(x,y)$ has a power spectral density $\Phi_{ss}(\omega_x, \omega_y)$, and that the current frame can be predicted up to some residual noise $n(x,y)$ of power spectral density $\Phi_{nn}(\omega_x, \omega_y)$ by translating the reconstructed, previous frame $r(x,y)$ by the true displacement (d_x, d_y). In the frequency domain, this signal model is

$$S(\omega_x, \omega_y) = e^{-j\omega_x \frac{d_x}{X} - j\omega_y \frac{d_y}{Y}} \cdot R(\omega_x, \omega_y) + N(\omega_x, \omega_y)\,, \tag{13}$$

where $S(\omega_x, \omega_y)$ and $N(\omega_x, \omega_y)$ are Fourier transforms (8) of $s(x,y)$ and $n(x,y)$, respectively. If we assume that the noise $n(x,y)$, the signal $s(x,y)$, and the displacement error

$$\begin{pmatrix} \Delta d_x \\ \Delta d_y \end{pmatrix} = \begin{pmatrix} d_x \\ d_y \end{pmatrix} - \begin{pmatrix} \hat{d}_x \\ \hat{d}_y \end{pmatrix} \tag{14}$$

are statistically independent, then the power spectral density of the prediction error e is

$$\begin{aligned}\Phi_{ee}(\omega_x, \omega_y) = {} & \Phi_{ss}(\omega_x, \omega_y) \\ & \cdot \left(1 + |F(\omega_x, \omega_y)|^2 \right. \\ & \left. - 2\Re\{F(\omega_x, \omega_y) P(\omega_x, \omega_y)\}\right) \\ & + \Phi_{nn}(\omega_x, \omega_y) |F(\omega_x, \omega_y)|^2\,,\end{aligned} \tag{15}$$

[23], where $\Re\{\cdot\}$ denotes the real part of a complex number, and $P(\omega_x, \omega_y)$ is the band-limited 2-D Fourier transform of the continuous probability density function (pdf) $p(\Delta d_x, \Delta d_y)$ of the displacement error $(\Delta d_x, \Delta d_y)$,

$$P(\omega_x, \omega_y) = \iint_{x,y} P(x,y) e^{-j\omega_x x/X - j\omega_y y/Y}\, dx\, dy \qquad \forall\ |\omega_x| < \pi, \qquad |\omega_y| < \pi\,. \tag{16}$$

Interestingly, frequency components of the displacement error pdf outside the baseband have no influence on the prediction error power spectrum. By differentiating (15) with respect to $F(\omega_x, \omega_y)$, it is readily shown that the mean squared prediction error is minimized at each frequency if $F(\omega_x, \omega_y)$ is a Wiener filter with frequency response

$$F(\omega_x, \omega_y) = P^*(\omega_x, \omega_y) \cdot \frac{\Phi_{ss}(\omega_x, \omega_y)}{\Phi_{ss}(\omega_x, \omega_y) + \Phi_{nn}(\omega_x, \omega_y)}\,. \tag{17}$$

The superscript * is used to denote complex conjugation. Wiener filter (17) can be interpreted to consist of two stages: one is a Wiener filter with respect to the noise $n(x,y)$; the other one is a Wiener filter that takes into account the inaccuracy of the displacement estimate.

Equation (15) allows us to study the influence of the displacement error pdf on the prediction error variance (1), which can be calculated on the basis of Parseval's relation

$$\sigma_e^2 = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \Phi_{ee}(\omega_x, \omega_y)\, d\omega_x\, d\omega_y\,. \tag{18}$$

We have evaluated the prediction error variance assuming an isotropic signal power spectrum

$$\Phi_{ss}(\omega_x, \omega_y) = \frac{2\pi\sigma_s^2}{\omega_0^2} \cdot \left(1 + \frac{\omega_x^2 + \omega_y^2}{\omega_0^2}\right)^{-\frac{3}{2}} \tag{19}$$

[23], a flat noise power spectrum

$$\Phi_{nn}(\omega_x, \omega_y) = \frac{\sigma_n^2}{4\pi^2}\,, \tag{20}$$

and an isotropic Gaussian displacement error pdf of variance $\sigma_{\Delta d}^2$

$$p(\Delta d_x, \Delta d_y) = \frac{1}{2\pi\sigma_{\Delta d}^2} \exp\left(-\frac{\Delta d_x^2 + \Delta d_y^2}{2\sigma_{\Delta d}^2}\right), \tag{21}$$

where σ_s^2 and σ_n^2 are signal and noise variances, respectively, and ω_0 has been set to correspond to a typical correlation of 0.93 between adjacent samples. Fig. 3 shows the influence of the displacement error variance $\sigma_{\Delta d}^2$ on the prediction error variance σ_e^2 for three variances σ_n^2 of noise contained in the

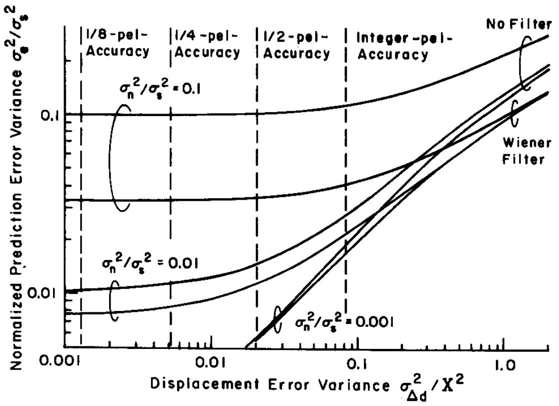


Fig. 3. Results of a model calculation showing the influence of motion-compensation accuracy on prediction error variance for noisy signals. The "filtering effect" is illustrated by the difference between the "no filter" and the "Wiener filter" curves. The vertical dashed lines indicate the minimum displacement error variances that can be achieved with the indicated motion-compensation accuracies in moving areas.

reconstructed picture $r(x,y)$. The curves compare a Wiener filter (17) to the case "no filter,"

$$F(\omega_x, \omega_y) = 1. \tag{22}$$

The following observations are important.

- Prediction error variance is generally decreased by more accurate motion-compensation.
- Beyond a certain "critical accuracy," the possibility of further improving prediction by more accurate motion-compensation is small.
- The critical point is at a high displacement error variance for high noise variance, and at a low displacement error variance for low noise variance.
- For low noise, the Wiener filter is more effective for less accurate motion-compensation than for accurate motion-compensation
- For high noise, the Wiener filter is more effective for accurate motion-compensation than for less accurate motion-compensation.
- For accurate motion-compensation, the potential of the Wiener filter increases with noise level.

Fig. 3 also indicates the minimum displacement error variance that can be achieved for a given motion-compensation accuracy in moving areas. Consider a perfect displacement estimator that always estimates the true displacement. Then, the displacement error $(\Delta d_x, \Delta d_y)$ is entirely due to rounding. In moving areas with sufficient variation of motion, the displacement error will be uniformly distributed between $\pm(1/2)\beta X$ and $\pm(1/2)\beta Y$, where $\beta = 1$ for integer-pel accuracy, $\beta = 1/2$ for 1/2-pel accuracy, etc. For a grid with balanced horizontal and vertical resolution, $X = Y$, the minimum displacement error variance in moving areas is

$$\sigma_{\Delta d}^2 = \frac{(\beta X)^2}{12}. \tag{23}$$

It turns out that the precise shape of the displacement error pdf has hardly any influence on the variance of the motion-compensated prediction error, σ_e^2, as long as the displacement error variance $\sigma_{\Delta d}^2$ does not change. A uniform pdf and a Gaussian pdf yield essentially the same variances σ_e^2.

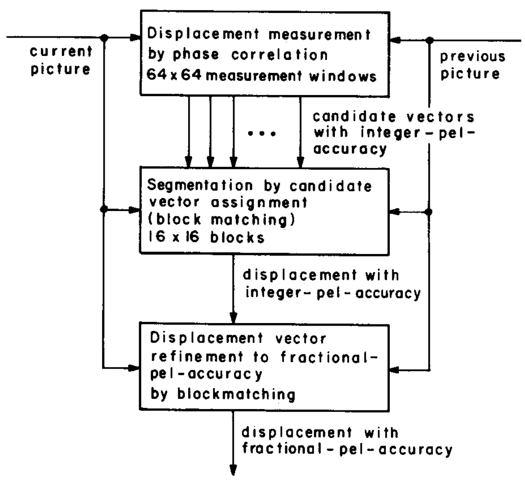

current picture
previous picture
Displacement measurement by phase correlation 64x64 measurement windows
candidate vectors with integer-pel-accuracy
Segmentation by candidate vector assignment (block matching) 16 x 16 blocks
displacement with integer-pel-accuracy
Displacement vector refinement to fractional-pel-accuracy by blockmatching
displacement with fractional-pel-accuracy

Fig. 4. Three-stage displacement estimator for fractional-pel accuracy.

IV. Displacement Estimation with Fractional-Pel Accuracy

After the model calculation in the previous section, we want to measure prediction error variance as a function of displacement error variance experimentally for actual pictures. In order to obtain meaningful results, it is very important to use a displacement estimator that gives both reliable and accurate results. If the displacement estimator itself is not accurate, we could hardly improve prediction by fractional-pel accuracy of MCP.

For our experiments, we have used a displacement estimation algorithm that consists of three stages (Fig. 4). It is based on the phase correlation technique, first proposed by Kuglin and Hines [25].

A. Displacement Measurement by Phase Correlation

Large overlapping measurement windows of size 64 x 64 are moved over the picture. Based on the Discrete Fourier Transform (DFT) [24] $S(\omega_x, \omega_y)$ of a block of the input signal $s(x,y)$, a frequency-weighted phase array

$$\Psi(\omega_x, \omega_y) = W(\omega_x, \omega_y) \cdot \frac{S(\omega_x, \omega_y) \cdot S_{-1}^*(\omega_x, \omega_y)}{|S(\omega_x, \omega_y) \cdot S_{-1}^*(\omega_x, \omega_y)|} \tag{24}$$

is calculated, where $S_{-1}(\omega_x, \omega_y)$ is the DFT of the corresponding block in the previous picture. Again, the superscript * is used to denote complex conjugation. The inverse DFT of the phase array $\Psi(\omega_x, \omega_y)$ is the so-called phase correlation surface; it can be shown to contain one impulse for each translatorily moving patch within the measurement window [26]. The size of each peak indicates the area (or rather the ac energy) that the patch covers within the window. Phase correlation can be interpreted as an estimate of the displacement histogram within the measurement window [27]. The weighting function $W(\omega_x, \omega_y)$ smoothes the phase correlation surface and suppresses spurious peaks in the case of noninteger displacements. We used a separable Kaiser window with $\alpha = 2.0$ [28]. The predominant peaks in the phase correlation surface are detected with integer-pel accuracy and passed on as "candidate displacements" to stage B of the estimator.

B. Segmentation by Candidate Vector Assignment

A candidate vector from the corresponding large phase correlation measurement window, that provides the best match for a smaller "block," is assigned to that block. The size of these blocks was 16 × 16 in our experiments. The vector assignment is done by block matching [1] using the accumulated magnitude of the displaced frame difference, $\sum |DFD|$, as a matching criterion. Unlike conventional block matching [1], only displacement vectors "suggested" as candidates by stage A are tested, so that the computational load is greatly reduced while the result is far more reliable. The 16 × 16 blocks do not overlap, hence each block is assigned a unique displacement vector.

C. Displacement Vector Refinement by Search

Stages A and B of the estimator consider only integer displacements. For a refinement to fractional-pel accuracy, a block matching search procedure starts with the vectors from stage B. The matching criterion is again $\sum |DFD|$, accumulated over a 16 × 16 block. The search procedure first considers all combinations of displacements by $\pm(1/2)X$ horizontally and $\pm(1/2)Y$ vertically. Then, starting from the 1/2 pel displacement with minimum $\sum |DFD|$, all 1/4-pel displacements are compared. The procedure is repeated with displacement step size 2^{-n} until the desired accuracy of the displacement estimate is obtained. Samples at fractional-pel positions needed in the displacement search were computed using the "sinc-interpolation" described in more detail in Section V.

A similar combination of stages A and B has been recently proposed by Thomas [26]. His report also presents an analysis of the properties of phase correlation, which need not be repeated here. We have augmented his two-stage procedure by the third stage for vector refinement. Thomas proposes to measure fractional-pel displacements already in the phase correlation stage A. While phase correlation has the potential to measure displacements with an error much smaller than the interval between samples, we found that there are problems for smoothly varying displacement vector fields which are very common in natural image sequences. If the phase correlation measurement window covers only one constant displacement, the phase correlation surface will be a shifted version of the inverse DFT of the weighting function $W(\omega_x, \omega y)$. If we would use no weighting, i.e., $W(\omega_x, \omega_y) = 1$, the "impulse" indicating the displacement would be a shifted, sampled sinc-function with a main-love width $2X$ horizontally and $2Y$ vertically [26]. The weighting function additionally widens this main lobe. If the displacement varies within the measurement window by small amounts, this will also have the effect of widening the peak. It is hard, if not impossible, to select a number of fractional-pel displacements out of such a smeared peak. Of course, we could consider all fractional-pel displacements under a peak as candidate vectors for stage B; this would, however, be computationally very expensive. For 1/8-pel accuracy, e.g., it would require 64 times as many candidate vectors to be tested in stage B than for integer-pel accuracy.

Fig. 5(d) shows a typical vector field obtained with the three-stage displacement estimator. Vector fields are smooth within homogeneously moving regions. The spilling of nonzero motion vectors into the static background region ("corona effect"), a typical problem for some other hierarchical schemes, is avoided. Note that zero-length vectors are not displayed in Fig. 5(d). With respect to variance of the displaced frame difference, the estimator outperforms other state-of-the-art schemes like hierarchical block matching [29] of differential techniques [30] for a given vector accuracy and block size.

V. Experimental Comparison of Various Spatial Prediction/Interpolation Filters

With fractional-pel accuracy displacements estimated by the three-stage procedure of the previous section, three different types of spatial prediction/interpolation filters $F(\omega_x, \omega_y)$ were compared experimentally at different motion-compensation accuracies.

1) Sinc-interpolation corresponds to the case $F(\omega_x, \omega_y) = 1$, or "no filter" in Fig. 3. Convolution with $f_2(x, y)$ (6) is omitted in this case. For a horizontally and vertically band-limited signal, the interpolation kernel (7) would ideally be

$$\tilde{f}_1(x, y) = \frac{\sin(\pi x/X) \cdot \sin(\pi y/Y)}{(\pi x/X) \cdot (\pi y/Y)}. \tag{25}$$

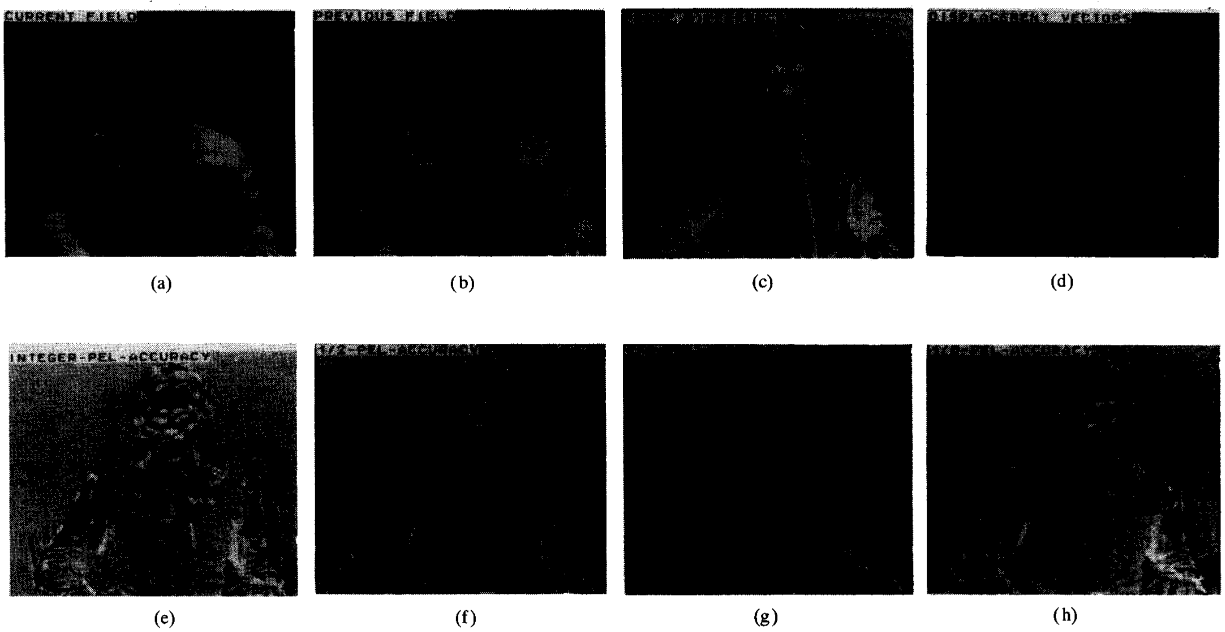
The sinc-interpolation kernel (25) has infinite extent. For a practical system, we have to approximate (25) by a finite impulse response. We have cascaded horizontal and vertical 2:1 interpolations with carefully designed 21-tap filters to calculate samples at fractional-pel positions. Interpolations by the factors 4:1 and 8:1 were performed by multiple passes through the 2:1 interpolators.

2) Bilinear interpolation uses the interpolation kernel

$$\tilde{f}_1(x, y) = \max\left\{0, 1 - \left|\frac{x}{X}\right|\right\} \cdot \max\left\{0, 1 - \left|\frac{y}{Y}\right|\right\}. \tag{26}$$

Again, convolution with $f_2(x, y)$ (6) is omitted. Sinc-interpolation and bilinear interpolation are identical for integer-pel accuracy of motion-compensation.

3) Wiener filters were computed separately for each estimation accuracy and each source signal. Equation (17) gives the Wiener filter as a function of signal and noise power spectra and displacement error pdf. Although this formulation is useful to understand MCP, it is not a useful formulation for filter design, since noise power spectrum and displacement error pdf are usually not known explicitly. We take an alternative approach here. Let us call $r_w(x, y)$ (7) the *compensated reconstructed previous frame* $c(x, y)$ when a sinc-interpolation kernel (25) is used in (7). Our task is to find the filter that best predicts $s(x, y)$ when applied to $c(x, y)$. It is a well-known



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 5. Picture pair TREVOR: (a) current picture, (b) previous picture, (c) frame difference without motion-compensation, (d) displacement vectors estimated by the three-stage procedure described in the text, (e) motion-compensated prediction error with integer-pel accuracy, (f) motion-compensated prediction error with 1/2-pel accuracy, (g) motion-compensated prediction error with 1/4-pel accuracy, (h) motion-compensated prediction error with 1/8-pel accuracy.

result from linear mean-square estimation [31] that the Wiener filter with frequency response

$$F(\omega_x, \omega_y) = \frac{\Phi_{sc}(\omega_x, \omega_y)}{\Phi_{cc}(\omega_x, \omega_y)} \tag{27}$$

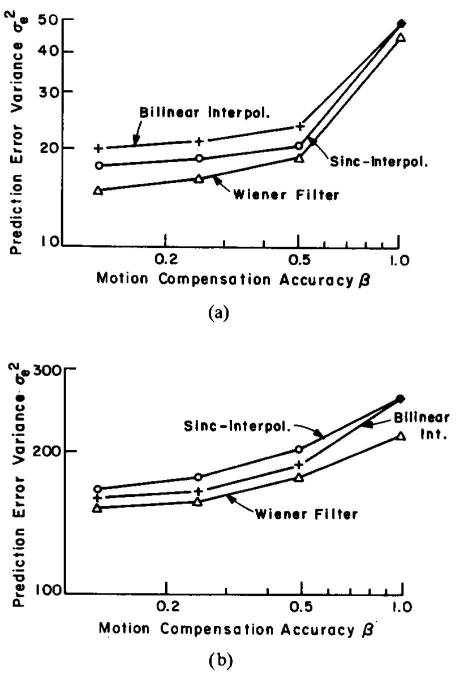
minimizes the mean squared prediction error. In (27), $\Phi_{sc}(\omega_x, \omega_y)$ is the cross spectrum between $s(x, y)$ and $c(x, y)$, and $\Phi_{cc}(\omega_x, \omega_y)$ is the power spectrum of $c(x, y)$. Both spectra can be measured directly. With the assumption that the previous reconstructed picture $r(x, y)$ does not differ from the previous original picture, thus neglecting coding noise, $\Phi_{sc}(\omega_x, \omega_y)$ and $\Phi_{cc}(\omega_x, \omega_y)$ have been estimated from the signal by averaging periodograms of 16 × 16 blocks [24]. This results in 16 × 16 samples of the Wiener filter frequency response and, after inverse DFT, in an impulse response with a 16 × 16 region of support. The Wiener filter impulse response typically decays very fast, and most of the 16 × 16 coefficients are very close to zero. The 16 × 16 impulse response was used as $f_2(x, y)$ in (6), and combined with the sinc-interpolation described as item 1) in this section.

The variance of the prediction error for the different spatial prediction/interpolation filters is shown in Figs. 6 and 7 as a function of motion-compensation accuracy. The prediction was based on the previous original picture rather than the previous reconstructed picture $r(x, y)$, thus neglecting noise introduced by encoding of the prediction error e (Fig. 1). We show results for two broadcast TV signals—ZOOM and VOITURE—and two videophone signals—TREVOR (Fig. 5) and MISS AMERICA. The source material is described in detail in the Appendix. For the videophone signals TREVOR and MISS AMERICA, only moving parts have been considered both for designing the Wiener filter and measuring the prediction error variance. Figs. 6 and 7 illustrate the following observations.

- Compared to integer-pel accuracy of MCP without filtering, prediction error variance can be reduced by more accurate motion-compensation and Wiener filtering by up to 5.2 dB for ZOOM, 2.3 dB for VOITURE, 1.8 dB for TREVOR, and 0.7 dB for MISS AMERICA.
- Except for ZOOM, bilinear interpolation is as good as, or better than, sinc-interpolation.
- For the broadcast TV signals ZOOM and VOITURE, MCP with 1/4-pel accuracy is certainly sufficiently accurate for a practical coder.
- For the videophone signals TREVOR and MISS AMERICA, 1/2-pel accuracy seems to be a desirable limit.
- The curves in Figs. 6 and 7 qualitatively correspond to the model curves in Fig. 3. The measurement results can be explained by the theory presented in Section III.

For the videophone signals and 1/2-pel accuracy of motion-compensation, we have simplified the Wiener filter to a separable filter that is suitable for a hardware realization. This filter uses the coefficients $(1/8, 6/8, 1/8)$ centered around the estimated motion trajectory, if the estimated displacement component is integer. If the estimated displacement requires a half-pel-shift, the coefficients $(1/16, 7/16, 7/16, 1/16)$ are used. This filter, denoted as $(1, 2, 7, 12, 7, 2, 1)$, is compared in Table I to bilinear interpolation and to Wiener filtering. The additional gains over bilinear interpolation are rather small.

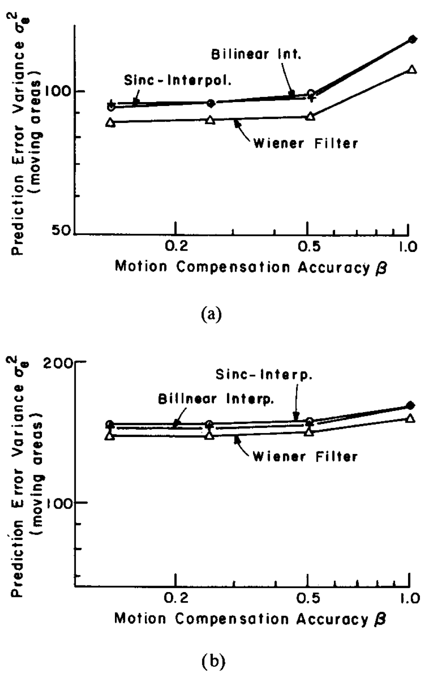
Fig 5(e)–(h) shows the motion-compensated prediction error $e(x, y)$ for TREVOR at different accuracies of MCP. For this example, sinc-interpolation was used. This series of



(a)

(b)

Fig. 6. Experimental comparison of different spatial prediction/interpolation filters for braodcast TV signals (a) ZOOM and (b) VOITURE. Scaling of the axis "motion-compensation accuracy β" is such that $\beta = 1$ corresponds to integer-pel accuracy, $\beta = 1/2$ corresponds to $1/2$-pel accuracy, etc.



(a)

(b)

Fig. 7. Experimental comparison of different spatial prediction/interpolation filters for videophone signals (a) TREVOR and (b) MISS AMERICA. Motion-compensated prediction error variance measured for moving areas only. Scaling of the axis "motion-compensation accuracy β" is such that $\beta = 1$ corresponds to integer-pel accuracy, $\beta = 1/2$ corresponds to $1/2$-pel accuracy, etc.

TABLE I

PREDICTION ERROR VARIANCE FOR MOVING AREAS, COMPARISON OF INTERPOLATION WITH THE FILTER (1, 2, 7, 12, 7, 2, 1) DESCRIBED IN THE TEXT WITH BILINEAR INTERPOLATION AND WITH WIENER FILTERING. MOTION-COMPENSATION IS PERFORMED WITH 1/2-PEL ACCURACY.

Signal	Bilinear interpolation	Filter (1, 2, 7, 12, 7, 2, 1)	Wiener filter
TREVOR	97.6	92.4	89.5
MISS AMERICA	147.5	146.5	142.5

pictures provides an intuitive explanation of the curves in Figs. 6 and 7. Once MCP with a certain critical accuracy has been applied, the motion-compensated prediction error contains mostly components that cannot be further reduced by more accurate motion-compensation. The signal model "constant displacement within a block" is not sufficient to describe all signal changes occurring from one frame to the next. There is a variety of effects that limit the efficiency of MCP, most of which can be discovered in Fig. 5(e)–(h). A meaningful displacement does not exist where background is uncovered. There can be spatial resolution changes due to zoom, varying distance between camera and object, or temporal integration of the camera target. The apparent brightness of a surface, in general, varies when any of the angles between surface normal, light source, and observer changes, or when shadows are cast on the surface. Problems also occur at object borders, where displacement is spatially rapidly varying, or with rotational movements. In addition to all this, there might be shortcomings of the camera that introduce noise and aliasing. All these effects contribute to the residual prediction error $e(x, y)$ that is encoded and transmitted in the motion-compensating hybrid coder. In order to further reduce this residue in the future, we will have to use much more elaborate models than those underlying MCP.

VI. CONCLUSION

This paper studies the effect of fractional-pel accuracy on the efficiency of a motion-compensating predictor in conjunction with various spatial interpolation/prediction filters. MCP has been analyzed by relating the prediction error power spectrum to the probability density function of the displacement error. Model calculations explain the "accuracy effect," the "filtering effect," and a "critical accuracy," beyond which the possibility of further improving prediction by more accurate motion-compensation is small.

The model calculations were confirmed experimentally by comparing sinc-interpolation, bilinear interpolation, and Wiener filtering for fractional-pel accuracy MCP. An improved algorithm for reliable displacement estimation with fractional-pel accuracy was used in the experiments. In three stages, it employs phase correlation followed by candidate vector assignment and a vector refinement to fractional-pel accuracy.

Signal components that do not obey the paradigm of translatory motion limit the performance of MCP. It was found that, for a motion-compensation block size of 16 × 16 and typical broadcast TV signals, 1/4-pel accuracy appears to be sufficient; while for videophone signals, 1/2-pel accuracy is desirable. For videophone signals with bilinear interpolation, the filtering effect is partly exploited in addition to the accuracy effect. A new separable filter, that gives a slightly better prediction for 1/2-pel accuracy, has been proposed.

Appendix

Picture Material Used for the Experiments

ZOOM: Two fields out of a broadcast TV sequence (DOLL40 kindly provided by Deutsche Thomson-Brandt), sampled at 13.5 MHz with line-interlace. The two fields are taken 40 ms apart. The processed window is 128 lines × 256 pels and shows a building and parts of a ship with rich detail. The movement is generated exclusively by camera zoom.

VOITURE: Two fields 20 ms apart from a broadcast TV sequence (VOITURE kindly provided by CCETT) sampled at 13.5 MHz with line-interlace. Mainly horizontal motion of rigid objects due to camera pan, motion of the car, and motion of the gate. Contains much uncovered background. Processed window: 256 lines × 512 pels.

TREVOR: A videophone sequence (TREVOR kindly provided by British Telecom Research Laboratories) has been converted to a format of 7.5 noninterlaced frames per second with 288 lines × 352 pels. Field numbers 10 and 11 of this sequence have been processed. The processed window is 256 lines × 256 pels. The picture pair contains motion up to 9 lines vertically and up to 6 pels horizontally.

MISS AMERICA: A videophone sequence which again has been converted to a format of 7.5 noninterlaced frames per second with 288 lines × 352 pels. Field numbers 5 and 6 of this sequence have been processed. The processed window is 256 lines × 256 pels. The picture pair represents moderate motion combined with a significant change in facial expression.

Acknowledgment

The author wishes to thank Prof. Dr.-Ing. H. G. Musmann, the staff of Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung, Universität Hannover, and his colleagues in the COST 211bis simulation subgroup for many helpful discussions. F. Joubert designed and simulated the Wiener filters.

References

[1] J. R. Jain and A. K. Jain, "Displacement measurement and its application in interframe image coding," *IEEE Trans. Commun.*, vol. COM-29, pp. 1799–1804, Dec. 1981.

[2] T. Ishiguro and K. Iinuma, "Television bandwidth compression by motion-compensated interframe coding," *IEEE Commun. Mag.*, pp. 24–30, Nov. 1982.

[3] T. Koga, A. Hirano, K. Iinuma, Y. Iijima, and T. Ishiguro, "A 1.5 Mb/s interframe codec with motion-compensation," in *Proc. Int. Conf. Commun.*, Boston, MA, 1983, pp. 1161–1165.

[4] T. Murakami, K. Asai, A. Itoh, and E. Yamazaki, "Interframe vector coding of color video signals," in *Proc. Int. Picture Coding Symp.*, Rennes, France, 1984.

[5] G. Kummerfeldt, F. May, and W. Wolf, "Coding television signals at 320 kbit/s and 64 kbit/s," *SPIE*, vol. 594, Image Coding, pp. 119–128, Dec. 85.

[6] S. Ericsson, "Fixed and adaptive predictors for hybrid predictive/transform coding," *IEEE Trans. Commun.*, vol. COM-33, pp. 1291–1302, Dec. 1985.

[7] M. J. Bage, "Interframe predictive coding of images using hybrid vector quantization," *IEEE Trans. Commun.*, vol. COM-34, pp. 411–415, Apr. 1986.

[8] H. Brusewitz and P. Weiss, "A video conference sytem at 384 kbit/s," in *Proc. Int. Picture Coding Symp.*, Tokyo, Japan, 1986.

[9] A. V. Brandt, "Motion estimation and subband coding using quadrature mirror filters," presented at EUSIPCO '86, The Hague, The Netherlands, Sept. 1986.

[10] H. Murakami, S. Matsumoto, Y. Hatori, and H. Yamamoto, "15/30 Mbit/s universal digital TV codec using a median adaptive predictive coding method," *IEEE Trans. Commun.*, vol. COM-35, pp. 637–645, June 1987.

[11] M. Kaneko, Y. Hatori, and A. Kloike, "Improvements of transform coding algorithm for motion-compensated interframe prediction errors—DCT/SQ Coding," *IEEE J. Select. Areas Commun.*, vol. SAC-5, pp. 1068–1078, Aug. 1987.

[12] P. Gerken and H. Schiller, "A low bit-rate image sequence coder combining a progressive DPCM on interleaved rasters with a hybrid DCT technique," *IEEE J. Select. Areas Commun.*, vol. SAC-5, pp. 1079–1089, Aug. 1987.

[13] Y. Kato, N. Mukawa, and S. Okubo, "A motion picture coding algorithm using adaptive DCT encoding based on coefficient power distribution classification," *IEEE J. Select. Areas Commun.*, vol. SAC-5, pp. 1090–1099, Aug. 1987.

[14] R. J. Moorhead, II, S. A. Rajala, and L. W. Cook, "Image sequence compression using a pel-recursive motion-compensated technique," *IEEE J. Select. Areas Commun.*, vol. SAC-5, pp. 1100–1114, Aug. 1987.

[15] T. Koga and M. Ohta, "Entropy coding for a hybrid scheme with motion-compensation in subprimary rate video transmission," *IEEE J. Select Areas Commun.*, vol. SAC-5, pp. 1166–1174, Aug. 1987.

[16] D. R. Walker and K. R. Rao, "Motion-compensated coder," *IEEE Trans. Commun.*, vol. COM-35, pp. 1171–1178, Nov. 1987.

[17] J. Biemond, B. P. Thieme, and D. E. Boekee, "Subband-coding of moving video using hierarchical motion estimation and vector quantization," in *Proc. Int. Workshop 64 kbit/s Coding Moving Video,* Hannover, Germany, June 1988.

[18] B. Girod and F. Joubert, "Motion-compensating prediciton with fractional pel accuracy for 64 kbit/s coding of moving video," in *Proc. Int. Workshop 64 kbit/s Coding Moving Video, Hannover, Germany, June 1988.*

[19] F. May, "Model based movement compensation and interpolation for ISDN videotelephony," presented at the 1988 IEEE Int. Symp. Circuits Syst. (ISCAS 88), Espoo, Finland, June 1988.

[20] H. Hashimoto, H. Watanabe, and Y. Suzuki, "A 64 kb/s video coding system and its performance," in *Proc. SPIE Conf. Visual Commun. Image Proc. '88,* Cambridge, MA, SPIE, vol. 1001, Nov. 1988, pp. 847–853.

[21] M. Gilge, "A high quality videophone coder using hierarchical motion estimation and structure coding of the prediction error," in *Proc. SPIE Conf. Visual Commun. Image Proc. '88,* Cambridge, MA, SPIE, vol. 1001, Nov. 1988, pp. 864–874.

[22] P. Strobach, "Tree-structured scene-adaptive coder," *IEEE Trans. Commun.*, vol. 38, pp. 477–486, Apr. 1990.

[23] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE J. Select. Areas Commun.*, vol. SAC-5, pp. 1140–1154, Aug. 1987.

[24] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1975.

[25] C. D. Kuglin, and D. C. Hines, "The phase correlation image alignment method," in *Proc. IEEE 1975 Int. Conf. Cybern. Soc.*, Sept. 1975, pp. 163–165.

[26] G. A. Thomas, "Television motion measurement for DATV and other applications," BBC Res. Dep. Rep. 1987/11, Sept. 1987.

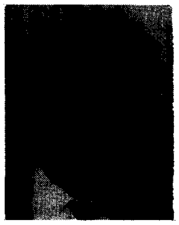
[27] B. Girod and D. Kuo, "Direct estimation of displacement histograms," in *Image Understanding and Machine Vision,* OSA 1989 Tech. Dig. Series, vol. 14, July 1989, pp. TuB3-1–TuB3-4.

[28] J. F. Kaiser, "Nonrecursive digital filter design using the I_0-*sinh* window function," in *Proc. 1974 Int. Symp. Circuits Syst.*, Apr. 1974, pp. 20–23.

[29] M. Bierling, "Displacement estimation by hierarchical block matching," in *Proc. SPIE Conf. Visual Commun. Image Proc. '88,* Cambridge, MA, SPIE, vol. 1001, Nov. 1988, pp. 942–951.

[30] ——, "A differential displacement estimation algorithm with improved stability," in *Proc. 2nd Int. Tech. Symp. Opt. Electro-Opt. Appl. Sci. Eng.*, SPIE Conf. B594 Image Coding, Cannes, France, Dec. 1985, pp. 170–174.

[31] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* New York: McGraw-Hill, 1965.



Bernd Girod (S'80–M'89) received the M.S. degree in electrical engineering from the Georgia Institute of Technology, in 1980 and the Doctoral degree "with highest honors" from the 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 Massachussetts Institute of Technology, Cambridge, MA, first as a Visiting Scientist with the Research Laboratory of Electronics, then as an Assistant Professor of Media Technology at the Media Laboratory. Since 1990, he is Professor of Computer Graphics and Technical Director of the Academy of Media Arts in Cologne, Germany. Additionally, he teaches at 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. His research interests include multidimensional signal processing, information theory, video signal compression, human and machine vision, sensory computing, computer graphics and animation, as well as interactive media. For several years, he has served as a consultant to companies and government agencies. He is also cofounder and Chief Scientist of Vivo Software Inc., Boston, MA.

Dr. Girod is the principal author of over 40 papers or book chapters in his field. He serves on the Editorial Boards of the journals IEEE TRANSACTIONS ON IMAGE PROCESSING, *Computer and Graphics, Visual Communication and Image Representation, and Image Communication.* He was chaired the 1990 SPIE conference on "Sensing and Reconstruction of Three-Dimensional Objects and Scenes" in Santa Clara, CA, and the 1993 German Multimedia Conference in Munich. He is a member of the German Informationstechnische Gesellschaft des VDE (ITG) and the German Gesellschaft für Informatik, serving on Fachausschuß 4.1, serving on the Image and Multidimensional Signal Processing Committee.