Analysis of Affine Motion-Compensated Prediction in Video Coding

Holger Meuel, and Jörn Ostermann

***Abstract*— Motion-compensated prediction is used in video coding standards like *High Efficiency Video Coding* (HEVC) as one key element of data compression. Commonly, a purely translational motion model is employed. In order to also cover non-translational motion types like rotation or scaling (zoom), e. g. contained in aerial video sequences such as captured from unmanned aerial vehicles (UAV), an affine motion model can be applied. In this work, a model for affine motion-compensated prediction in video coding is derived. Using the rate-distortion theory and the displacement estimation error caused by inaccurate affine motion parameter estimation, the minimum required bit rate for encoding the prediction error is determined. In this model, the affine transformation parameters are assumed to be affected by statistically independent estimation errors, which all follow a zero-mean Gaussian distributed probability density function (pdf). The joint pdf of the estimation errors is derived and transformed into the pdfof the location-dependent displacement estimation error in the image. The latter is related to the minimum required bit rate for encoding the prediction error. Similar to the derivations of the fully affine motion model, a four-parameter simplified affine model is investigated. Both models are of particular interest since they are considered for the upcoming video coding standard *Versatile Video Coding* (VVC) succeeding HEVC. Both models provide valuable information about the minimum bit rate for encoding the prediction error as a function of affine estimation accuracies.**

***Index Terms*— Video coding, (simplified) affine motion-compensated prediction (MCP), rate-distortion theory, Versatile Video Coding (VVC).**

I. Introduction

MODERN hybrid video coding standards like *Advanced Video Coding* (AVC) [1], or *High Efficiency Video Coding* (HEVC) [2] provide very good video compression capabilities for daily life applications like Digital Video Broadcasting (DVB) [3]. Furthermore, video on demand (VOD) applications, e. g. like Netflix or Amazon Prime Video, and also internet video applications like Youtube depend on high video compression performance. However, video compression standards like HEVC are natively optimized for the compression of video sequences as produced by commercial movie production studios or home-brew videos such as captured with a smartphone, camcorder or other digital movie cameras. They reduce the redundancy contained in a video sequence by a combination of motion-compensated prediction (MCP), transform coding with quantization, both typically realized in a *differential pulse-code modulation* (DPCM) loop, and entropy coding [3]. MCP exploits that most parts of one video image (further on referred to as *frame*) reoccur in preceding or subsequent frames of the sequence. Instead of a pixel-wise representation of a certain, typically rectangular, image part (called *block*), only a displacement vector to a similar image block is stored (motion vector). For the most often used lossy coding schemes, the remaining pixel-wise prediction error is transformed using a decorrelating transform. Typically, a *discrete cosine transform* (DCT) is applied and the resulting coefficients are quantized afterwards. The motion information, the quantized transform coefficients as well as additional signaling data needed for video decoding (e. g. video dimensions, frame rate, block partitioning, etc.) are entropy encoded, e. g. by using a *context-adaptive binary arithmetic coding* (CABAC). For the first frame of a video sequence, which is intrinsically new, or blocks, for which no appropriate candidate for motion-compensated prediction is found, *intra-frame coding* or just *intra coding* can be applied as an alternative. Intra coding uses only the current frame and thus requires no other frames. In either case, a rate-distortion optimization (RDO) is used to test several encoding possibilities with different block sizes, partitioning as well as coding modes and the one which provides the best bit rate with respect to the introduced distortion is selected for final coding.

A. Motion-Compensated Prediction

As mentioned above, one of the key elements for data compression in modern hybrid video coding standards is motion-compensated prediction (MCP). Since for video sequences captured at typical frame rates between 24 and 60 frames per second (fps) the same content is visible in many frames, the coding efficiency using inter-frame coding with MCP is much higher compared to that of intra-frame coding. More specific, MCP does not attempt to describe the real motion of a block, but rather searches for the corresponding block with the highest similarity, i. e. with the lowest distortion, typically measured as *mean squared error* (MSE) or *sum of absolute differences* (SAD). For a highly accurate prediction, the prediction error is small (or optimally zero) and the entropy of the prediction error is smaller than for an inaccurate

Manuscript received December 11, 2019; revised March 24, 2020 and May 8, 2020; accepted June 1, 2020. Date of publication June 17, 2020; date of current version July 13, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sérgio De Faria. *(Corresponding author: Holger Meuel.)*

The authors are with the Institut für Informationsverarbeitung, Leibniz Universität Hannover, 30167 Hannover, Germany (e-mail: meuel@tnt.uni-hannover.de).

Digital Object Identifier 10.1109/TIP.2020.3001734

prediction. Consequently, also the minimum required bit rate for encoding the prediction error depends on the accuracy of the motion estimation, which can be specified by the variance of the displacement estimation error. The minimum bit rate of the prediction error of motion-compensated prediction as a function of the variance of the displacement estimation error was analyzed by Girod already in 1987 [4]. In his work he showed that "the spatial power spectrum of the motion-compensated prediction error can be calculated from the signal power spectrum and the displacement estimation error probability density function (pdf)" [4]. Finally, he related the power spectrum of the motion-compensated prediction error to the minimum bit rate for encoding the prediction error by applying the rate-distortion theory. His derivations are employed again in this work. In contrast to the work of Girod, a non-isotropic autocorrelation function of the video signal is assumed in this work based on measurements. Moreover, Girod assumed that the displacement estimation errors Δx and Δy in x- and y-direction are uncorrelated and location-independent which only holds true for translational motion. Translational motion is relatively easy to estimate and describes most of the block motion for general videos sufficiently accurate. Consequently, Girod modeled the displacement estimation error for translational motion with two degrees of freedom. Such a motion model was employed in video coding standards like H.261 [5], MPEG-1 [1] [6], MPEG-2 [7], H.263 [8], AVC [1], and HEVC [2].

For video sequences with distinct global motion, affine *global motion compensation* (GMC) was introduced in MPG-4 *Advanced Simple Profile (*MPEG-4 ASP*)* [9], which can also cover rotation, scaling (i. e. zooming) and shearing. Since the coding efficiency gains of GMC stayed behind the expectations for general video coding for natural scenes without prevalent global motion, GMC was removed from the MPEG-4 ASP successor AVC again and replaced by an improved *motion vector prediction* (MVP). With upcoming small and relatively cheap *unmanned aerial vehicles* (UAVs) like multicopters, aerial video sequences with distinct global motion that cannot be covered by a purely translational motion model, become increasingly important. The importance of such sequences is also reflected in recent test sets, which contain more aerial video sequences than traditional video test sets, e. g. as used for the standardization of AVC or HEVC [10]–[13].

To improve the processing of such higher-order global motions, the ITU-T/ISO/IEC [2] *Joint Video Exploration Team* (JVET) incorporated a simplified 4-parameter affine motion model [14] (also referred to as *similarity* with four degrees of freedom, e. g. by Hartley and Zissermann [15]) into the experimental software *Joint Exploration Model (*JEM*)* [16] of the upcoming video coding standard *Versatile Video Coding* (VVC) again [17]. In contrast to MPEG-4 ASP, it operates on a block-level. Later, JVET additionally integrated a fully affine motion model with 6 degrees of freedom into the reference software VVC *Test Model* (VTM) [18]–[21]. Affine motion compensation is also part of the video codec (coder-decoder) *AOMedia Video* (AV1) [22], [23]. First investigations on the common test set [24] (containing no sequences consisting of distinct motion which cannot be covered by a purely translational model) show coding efficiency gains of up to 1.35 % [25], [26]. Larger gains of more than 20 % can be expected for sequences containing more higher-order motions [14], [20]. In [27], interweaved prediction is proposed to further enhance the coding efficiency. In that context, a theoretical analysis is carried out for the influence of interweaved prediction on the expected prediction error distribution within the prediction sub-blocks, and it is shown that the prediction error is decreased by interweaved prediction.

In this work, a theoretical model of the rate-distortion optimized bit rate for encoding the prediction error using affine (global) motion-compensated prediction is presented. For an affine motion model, particularly the assumption of Girod [4] of uncorrelated displacement estimation errors $\Delta x'$ and $\Delta y'$ (in the original work called Δx and Δy) in x- and y-direction cannot be applied for non-translational motion. Thus, in this work, the rate-distortion function for video coding using affine motion compensation is derived by extending the work of Girod [4] towards affine motion compensation and correlated displacement estimation errors $\Delta x'$ and $\Delta y'$. For this purpose the displacement estimation error during motion estimation is modeled and the bit rate after application of the rate-distortion theory is obtained, especially considering the power spectral density of modern high-resolution video sequences (Section II). It is noteworthy that the results of the derivations hold true for block-based as well as global motion compensation.

B. Contributions and Organization

The contribution of this work is the analysis of motion-compensated prediction using an affine motion model. Two different affine motion models are investigated, a fully one with 6 degrees of freedom and a simplified one with only 4 degrees of freedom.

For a fully affine motion model (with six degrees of freedom), the prediction error after motion compensation as a function of the affine transformation parameter accuracy is analytically derived. The affine parameters are assumed to be independently estimated and, as a worst-case assumption, independently perturbed by zero-mean Gaussian noise. Using the rate-distortion theory [28], the minimum required bit rate for encoding the prediction error is derived. More specifically, due to the assumptions as mentioned above, the supremum of the minimum required prediction error bit rate is derived.

Similar considerations are made for a simplified affine motion model with only four degrees of freedom (rotation, scaling, translation). Since the assumption of independently estimated affine transformation parameters cannot be met for the simplified model, the inter-correlation between the estimated parameters has to be specifically considered. Both models are investigated in the course of the standardization of VVC.

[1] MPEG : *Moving Picture Experts Group*.

[2] ITU-T: *International Telecommunication Union – Telecommunication Standardization Sector*; ISO: *International Organization for Standardization*; IEC: *International Electrotechnical Commission*.

The derivations for the fully affine model are based on [29], [30] and for the simplified affine model on [30]. In this work, all results are presented in a unified notation, related to each other, and thoroughly discussed [30]. Both models are valid for motion-compensated prediction applied on block-level or on entire frames as in the special case of global motion compensation.

In addition to the above derivations, the systematical error is modeled for the case that a purely translational motion model is employed for sequences containing non-translational affine motion. This systematical error is further related to the findings of the affine parameter estimation errors [30].

An exhaustive experimental validation of the findings is further presented and discussed in detail [30].

The remainder of this paper is organized as follows: in Section II, the efficiency of motion-compensated prediction is analyzed for a fully as well as for a simplified affine motion model and compared to the efficiency of a purely translational motion model using the example of aerial sequences containing distinct global motions. Experimental results are presented and discussed in Section III: the model from Section II is experimentally validated in Section III-A by measurements of the prediction error bit rate for inaccurate affine motion estimation. Operational rate-distortion diagrams for real-world sequences encoded with and without affine motion-compensated prediction are presented in Section III-B. Section IV summarizes and concludes this work.

II. Rate-Distortion Theory for Affine Motion Compensation in Video Coding

The largest contribution to the overall data rate of an encoded video stream in hybrid video coding is due to the encoding of the prediction error [31]. Thus, Bernd Girod modeled the minimum required bit rate for encoding the prediction error as a function of the motion estimation accuracy in his early work from 1987 [4]. In his work, Girod modeled the bit rate for a translational motion model and thus only for uncorrelated displacement estimation errors $\Delta x'$ and $\Delta y'$. With upcoming new application scenarios with video sequences containing distinct global and non-translational motion like aerial videos, it is beneficial to consider additional—non purely translational—motion models [14], [20], [21] as currently applied in the upcoming video coding standards *Versatile Video Coding* (VVC) [32] and AV1 [22], [23].

In this section an efficiency analysis of motion-compensated prediction is performed for a fully affine model [18], [19] with six degrees of freedom (Section II-A) as well as for a simplified affine motion model [14] (Section II-B). Both motion models currently are designated to be part of VVC [32].

To model the minimum required bit rate for encoding the prediction error, two different influences have to be distinguished. On the one hand, the model error itself has to be considered. The model error describes motions contained in the scene which cannot be covered by the selected motion model. On the other hand, the estimation error of the motion estimation itself has to be considered. The estimation error of course depends on the specific implementation and restrictions like motion vector accuracy in common hybrid video coding—as analyzed in [4]. Both aspects will be considered in this work. As for the rate-distortion analysis the source of the perturbations does not matter, the derivations for both are the same and thus are conducted only once. Similar as in [4], the model covers the minimum required bit rate for encoding the prediction error without any signaling. The latter may additionally account to a non-negligible bit rate. Parts of this section including the derivations for the fully and simplified affine models have been published in [29], [30].

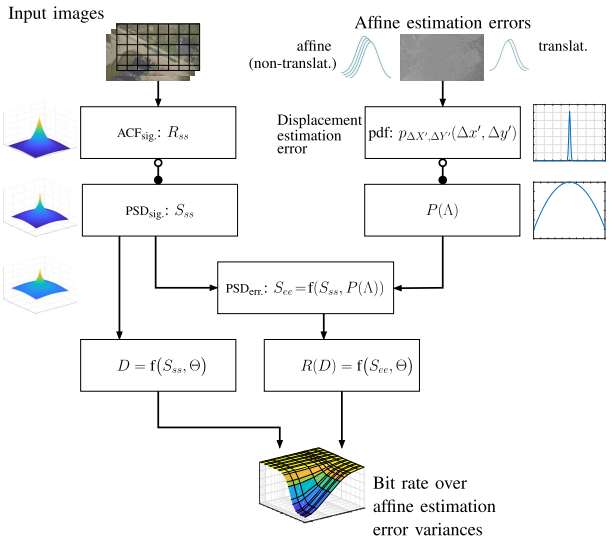


Fig. 1. Flowchart of the analysis.

A. Efficiency Analysis of Fully Affine Motion Compensation

The overview flow diagram in Fig. 1 illustrates the connections between the different components of the analysis within this section. The analysis is based on [4], although significant modifications have been introduced as explained in the next subsections.

The working steps are structured as follows:

- First, the affine motion and the error model as used for further derivations are introduced (Section II-A.1).
- Second, the 2D probability density function (pdf) $p_{\Delta X',\Delta Y'}(\Delta x', \Delta y')$ of the displacement estimation errors in x- ($\Delta x'$) and y-direction ($\Delta y'$) is derived (right part in Fig. 1). Here, $\Delta X'$ and $\Delta Y'$ denote the random processes generating $\Delta x'$ and $\Delta y'$. The Fourier transform of $p_{\Delta X',\Delta Y'}(\Delta x', \Delta y')$ is $P(\Lambda)$, which will be used for subsequent derivations as proposed by Girod [4]. Λ here abbreviates the two-dimensional (2D) spatial frequency vector $\Lambda := (\omega_x, \omega_y)$ for reasons of clarity (Sections II-A.2 and for the simplified affine model II-B.1).
- In a third step, the autocorrelation function (ACF) $R_{ss}(\Delta x', \Delta y')$ is modeled for typical input video sequences. The modeling is performed similar to that from O'Neal [33] and Girod [4] but was slightly modified in order to model also non-isotropic autocorrelation functions. According to the Wiener-Khinchin theorem,

the power spectral density (PSD) of the signal $S_{ss}(\Lambda)$ is the Fourier transform of this autocorrelation function $R_{ss}(\Delta x', \Delta y')$ (left part in Fig. 1, Section II-A.3).

- Combining the PSDof the signal $S_{ss}(\Lambda)$ and the Fourier transform of the probability density function of the displacement estimation error $P(\Lambda)$ by exploiting the findings from Girod [4], the PSDof the prediction error $S_{ee}(\Lambda)$ is derived (middle in Fig. 1, Section II-A.4).
- In the last step, the rate-distortion theory is applied to derive a distortion D and the corresponding bit rate $R(D)$ of the prediction error signal as proposed by Girod [4] (lower part in Fig. 1, Section II-A.5).
- The rate-distortion analysis of affine motion-compensated prediction is performed using real video signals for the fully affine (global) motion-compensated prediction in Section II-A.6 and for the simplified affine global motion-compensated prediction in Section II-B.2.
  First, in Section II-A.6.a, the affine parameter estimation error variances are determined for a real-world implementation. Based on the measurement, the probability density function of the displacement estimation error is calculated. Afterwards, the maximum gain which can be achieved by affine motion-compensated prediction instead of purely translational motion-compensated prediction is derived. Finally in this subsection, non-translational affine motions contained in representative camera-captured aerial video sequences were measured and related to the estimation error variances.
  Second, in Section II-A.6.b, the autocorrelation functions of real video sequences are measured. From the results, a mean power spectral density is derived. Third, in Section II-A.6.c, the rate-distortion theory is finally applied to determine the minimum required bit rate for encoding the prediction error.
- In Section II-A.7 finally conclusions are drawn for the fully affine motion-compensated prediction.

1) Affine Motion and Error Model: Assuming a fully affine motion model with six degrees of freedom, the x- and y-coordinates x' and y' in the source frame can be computed from the six affine parameters a_{ij} with $i=\{1,2\}$, $j=\{1,2,3\}$ and the coordinate $(x, y)^\top$ in the current (destination) frame in component notation by backwards prediction:

$$x' = a_{11}\cdot x + a_{12}\cdot y + a_{13}\,;\;\; y' = a_{21}\cdot x + a_{22}\cdot y + a_{23}\,. \quad (1)$$

The parameters a_{13} and a_{23} describe the translational part of a motion, whereas the parameters a_{11}, a_{12}, a_{21}, a_{22} express the rotation, scaling and shearing, respectively. It is assumed that each parameter a_{ij} is perturbed (indicated by $\hat{\cdot}$) by an independent error term e_{ij}, caused by inaccurate parameter estimation. Consequently, the perturbed coordinates $\hat{x}'$ and $\hat{y}'$ can be expressed as $\hat{x}' = \hat{a}_{11}x + \hat{a}_{12}y + \hat{a}_{13}$ and $\hat{y}' = \hat{a}_{21}x + \hat{a}_{22}y + \hat{a}_{23}$, leading to displacement estimation errors $\Delta x'$ and $\Delta y'$ (in pixel, further on referred to as pel) in horizontal and vertical direction of:

$$\begin{aligned}\Delta x' = \hat{x}' - x' &= \underbrace{(\hat{a}_{11} - a_{11})}_{e_{11}}\cdot x + \underbrace{(\hat{a}_{12} - a_{12})}_{e_{12}}\cdot y + \underbrace{(\hat{a}_{13} - a_{13})}_{e_{13}} \\ &= e_{11}\cdot x + e_{12}\cdot y + e_{13} \end{aligned} \quad (2)$$

$$\Delta y' = e_{21}\cdot x + e_{22}\cdot y + e_{23}. \quad (3)$$

2) Probability Density Function (pdf) of the Displacement Estimation Error: With the assumption that each error term e_{ij} is zero-mean Gaussian distributed, the probability density functions (pdfs) $p(e_{ij})$ of the error terms e_{ij} are

$$p(e_{ij}) = \frac{1}{\sqrt{2\pi\,\sigma_{e_{ij}}^2}}\cdot\exp\left(-\frac{e_{ij}^2}{2\sigma_{e_{ij}}^2}\right) \quad (4)$$

with $i = \{1,2\}$, $j = \{1,2,3\}$ and the variances $\sigma_{e_{ij}}^2$ of the error terms. For statistically independent variables the joint pdf $p_{E_{11},\ldots,E_{23}}(e_{11},\ldots,e_{23})$ for the random variables $E_{11},\ldots,E_{23}$ generating the observations $e_{11},\ldots,e_{23}$ is:

$$p_{E_{11},\ldots,E_{23}}(e_{11},\ldots,e_{23}) = p(e_{11})\cdot\;\ldots\;\cdot p(e_{23}). \quad (5)$$

To convert the pdf $p_{E_{11},\ldots,E_{23}}(e_{11},\ldots,e_{23})$ to the desired pdf $p_{\Delta X',\Delta Y'}(\Delta x', \Delta y')$ with the random processes $\Delta X'$, $\Delta Y'$ generating the resulting displacement estimation errors $\Delta x'$ and $\Delta y'$ as caused by affine parameter estimation errors, the transformation theorem for pdfs is used ([34], [35]):

$$\begin{aligned} p_{\mathscr{Y}_1,\ldots,\mathscr{Y}_M}(\mathscr{y}_1,\ldots,\mathscr{y}_M) = &\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} p_{\mathscr{X}_1,\ldots,\mathscr{X}_N}(\xi_1,\ldots,\xi_N) \\ &\cdot\prod_{m=1}^{M}\delta\big(\mathscr{y}_m - g_m(\xi_1,\ldots,\xi_N)\big)\mathrm{d}\xi_1\ldots\mathrm{d}\xi_N \end{aligned} \quad (6)$$

with $\delta(\cdot)$ denoting the Dirac delta function, $g_1,\ldots,g_M$ being functions $\mathscr{y}_1 = g_1(\mathscr{x}_1,\ldots,\mathscr{x}_N)$, ..., $\mathscr{y}_M = g_M(\mathscr{x}_1,\ldots,\mathscr{x}_N)$, $\mathscr{X}_1,\ldots,\mathscr{X}_N$ and $\mathscr{Y}_1,\ldots,\mathscr{Y}_M$ representing random processes and $p_{\mathscr{Y}_1,\ldots,\mathscr{Y}_M}(\mathscr{y}_1,\ldots,\mathscr{y}_M)$ being the joint pdf. With (2) and (3) this yields

$$\begin{aligned} &p_{\Delta X',\Delta Y'}\left(\Delta x', \Delta y'|x, y\right) \\ &= \int_{\mathbf{R}^6} p_{E_{11},\ldots,E_{23}}(e_{11},\ldots,e_{23}) \\ &\quad\cdot\delta\big(\Delta x' - (xe_{11} + ye_{12} + e_{13})\big) \\ &\quad\cdot\delta\big(\Delta y' - (xe_{21} + ye_{22} + e_{23})\big)\;\mathrm{d}e_{11}\ldots\mathrm{d}e_{23} \end{aligned} \quad (7)$$

with a dependency on the location coordinates x and y in the current frame. By using the properties of the delta function and substituting e_{13} and e_{23}, the integrals

$$\begin{aligned} &p_{\Delta X',\Delta Y'}\left(\Delta x', \Delta y'|x, y\right) \\ &= \int_{\mathbf{R}^4} p_{E_{11},\ldots,E_{22}}(e_{11}, e_{12}, \Delta x' - xe_{11} - ye_{12}, e_{21}, e_{22} \\ &\quad \Delta y' - xe_{21} - ye_{22})\;\mathrm{d}e_{11}\mathrm{d}e_{12}\mathrm{d}e_{21}\mathrm{d}e_{22} \end{aligned} \quad (8)$$

are solved. Exploiting the statistical independence from (5), the integrands are separated, which leads to

$$\begin{aligned} &p_{\Delta X',\Delta Y'}\left(\Delta x', \Delta y'|x, y\right) \\ &= \int_{\mathbf{R}^2} p_{E_{11},E_{12},E_{13}}(e_{11}, e_{12}, \Delta x' - xe_{11} - ye_{12})\,\mathrm{d}e_{11}\mathrm{d}e_{12} \\ &\quad\cdot\int_{\mathbf{R}^2} p_{E_{21},E_{22},E_{23}}(e_{21}, e_{22}, \Delta y' - xe_{21} - ye_{22})\,\mathrm{d}e_{21}\mathrm{d}e_{22}\,. \end{aligned} \quad (9)$$

For simplicity, (9) is separated into its x- and y-components and the following derivation is presented for the x-component

only. The y-component can be calculated accordingly. From (9) with (4) the pdf of $\Delta x'$ is determined:

$$\begin{aligned}
&p_{\Delta X'}(\Delta x'|x,y)\\
&= \int_{\mathbf{R}^2} p_{E_{11},E_{12},E_{13}}(e_{11},e_{12},\Delta x' - x e_{11} - y e_{12})\,\mathrm{d}e_{11}\mathrm{d}e_{12}\\
&= \underbrace{\frac{1}{\sqrt{2\pi\sigma_{e_{11}}^2}}\cdot\frac{1}{\sqrt{2\pi\sigma_{e_{12}}^2}}\cdot\frac{1}{\sqrt{2\pi\sigma_{e_{13}}^2}}}_{A}\\
&\quad\cdot\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(-\frac{e_{11}^2}{2\sigma_{e_{11}}^2}\right)\cdot\exp\left(-\frac{e_{12}^2}{2\sigma_{e_{12}}^2}\right)\\
&\quad\cdot\exp\left(-\frac{(\Delta x' - x e_{11} - y e_{12})^2}{2\sigma_{e_{13}}^2}\right)\mathrm{d}e_{11}\mathrm{d}e_{12}\\
&= A\cdot\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\Bigg(-\frac{1}{2\sigma_{e_{11}}^2\sigma_{e_{12}}^2\sigma_{e_{13}}^2}\\
&\quad\cdot\Big[\sigma_{e_{12}}^2\sigma_{e_{13}}^2\ e_{11}^2+\sigma_{e_{11}}^2\sigma_{e_{13}}^2\ e_{12}^2\\
&\quad+\sigma_{e_{11}}^2\sigma_{e_{12}}^2\left(\Delta x' - x e_{11} - y e_{12}\right)^2\Big]\Bigg)\mathrm{d}e_{11}\mathrm{d}e_{12}. \qquad (10)
\end{aligned}$$

Integration results in

$$\begin{aligned}
p_{\Delta X'}\left(\Delta x'|x,y\right) = &\frac{1}{\sqrt{2\pi\left(\sigma_{e_{11}}^2x^2+\sigma_{e_{12}}^2y^2+\sigma_{e_{13}}^2\right)}}\\
&\cdot\exp\left(-\frac{\Delta x'^2}{2\cdot\left(\sigma_{e_{11}}^2x^2+\sigma_{e_{12}}^2y^2+\sigma_{e_{13}}^2\right)}\right). \qquad (11)
\end{aligned}$$

After calculating the y-component accordingly, the resulting displacement estimation error pdfis obtained as

$$\begin{aligned}
&p_{\Delta X',\Delta Y'}(\Delta x',\Delta y'|x,y)\\
&\quad=\frac{1}{2\pi\sigma_{\Delta x'}\sigma_{\Delta y'}}\cdot\exp\left(-\frac{\Delta x'^2}{2\sigma_{\Delta x'}^2}\right)\cdot\exp\left(-\frac{\Delta y'^2}{2\sigma_{\Delta y'}^2}\right) \qquad (12)
\end{aligned}$$

$$\text{with } \sigma_{\Delta x'}^2=\sigma_{e_{11}}^2x^2+\sigma_{e_{12}}^2y^2+\sigma_{e_{13}}^2 \qquad (13)$$

$$\text{and } \sigma_{\Delta y'}^2=\sigma_{e_{21}}^2x^2+\sigma_{e_{22}}^2y^2+\sigma_{e_{23}}^2. \qquad (14)$$

It is obvious that the variances $\sigma_{\Delta x'}^2$ and $\sigma_{\Delta y'}^2$ depend on the location in the frame. For simplicity $p_{\Delta X',\Delta Y'}(\Delta x',\Delta y'|x,y)$ is abbreviated as $p_{\Delta X',\Delta Y'}(\Delta x',\Delta y')$ further on and accordingly.

3) Power Spectral Density of the Signal: The power spectral density $S_{ss}(\omega_x,\omega_y)$ is modeled according to O'Neal and Girod [4], [33]. There it was assumed that the statistics of each frame of the video sequence can be represented by the isotropic autocorrelation function

$$\begin{aligned}
R_{ss,\mathrm{iso}}(\Delta x',\Delta y') &= E\left[s(x',y')\cdot s(x'-\Delta x',y'-\Delta y')\right]\\
&=\exp\left(-\alpha\sqrt{\Delta x'^2+\Delta y'^2}\right) \qquad (15)
\end{aligned}$$

with $s(x',y')$ representing the signal at position (x',y') and accordingly, $\Delta x'$ and $\Delta y'$ denoting the shift in x- and y-direction, respectively, and $E[\cdot]$ representing the expectation value. Based on measurements, in this work the autocorrelation function is assumed to be non-isotropic, leading to the general form

$$R_{ss}(\Delta x',\Delta y')=\exp\left(-\sqrt{\alpha_x^2\Delta x'^2+\alpha_y^2\Delta y'^2}\right). \qquad (16)$$

The exponential drop rates α_x and α_y in x- and y-direction can be determined as the negative logarithm of the correlations between horizontally and vertically adjacent pels $\alpha_x=-\ln(\rho_{ss,x})$ and $\alpha_y=-\ln(\rho_{ss,y})$ [33]. For this, the autocorrelation coefficients [35], [36] $\rho_{ss,x}$, $\rho_{ss,y}$ are calculated line- and column-wise, respectively. The power spectral density $S_{ss}(\Lambda)$ now is the Fourier transform of (16) (Wiener-Khinchin theorem).

4) Power Spectral Density of the Displacement Estimation Error: To derive the bit rate for encoding the prediction error in motion-compensated video coding, the findings from Girod are used [4]. He related the displacement estimation error pdf $p_{\Delta X',\Delta Y'}(\Delta x',\Delta y')$ to the prediction error e as follows: given a displacement estimation error pdf $p_{\Delta X',\Delta Y'}(\Delta x',\Delta y')$, the power spectral density of the prediction error

$$S_{ee}(\Lambda)=2\,S_{ss}(\Lambda)\left[1-\mathrm{Re}\{P(\Lambda)\}\right]+\Theta \qquad (17)$$

is obtained [4], where $S_{ss}(\Lambda)$ denotes the power spectral density of the video signal s, Λ the two-dimensional (2D) spatial frequency vector $\Lambda:=(\omega_x,\omega_y)$, $P(\Lambda)$ the 2D Fourier transform of the probability density function of the displacement estimation error, $\mathrm{Re}\{P(\Lambda)\}$ the real part of $P(\Lambda)$, and Θ a parameter that generates the rate-distortion function $R(D)$ (see next subsection) by taking on all positive real values ([4], Equation (28)). By variation of Θ the distortion and the corresponding rate for encoding the prediction error are determined, whereby one specific Θ yields one distinct distortion and a corresponding rate.

5) Rate-Distortion Function: Applying the rate-distortion theory [28] finally results in the minimum required bit rate for encoding the prediction error. The distortion D as well as the corresponding minimum bit rate $R(D)$ are derived from the rate-distortion function for a given mean-squared error (Equations (19), (20) in [4], and [28]):

$$D=\frac{1}{4\pi^2}\iint_{\Lambda}\min\left[\Theta,S_{ss}(\Lambda)\right]\mathrm{d}\Lambda \qquad (18)$$

$$R(D)=\frac{1}{8\pi^2}\iint_{\substack{\Lambda:\ \left(S_{ss}(\Lambda)>\Theta\\ \text{and } S_{ee}(\Lambda)>\Theta\right)}}\log_2\left[\frac{S_{ee}(\Lambda)}{\Theta}\right]\mathrm{d}\Lambda\ \text{ bit}. \qquad (19)$$

Both, Equations (18) and (19), are connected by the generating function Θ, which was also used in Equation (17). As explained above,"generating function" here means that an arbitrary positive real value can be selected. Then, one specific distortion D can be calculated for the selected value of Θ (and of course as a function of the power spectral density of the video signal $S_{\mathrm{ss}}(\Lambda)$). The same value of Θ that was used for one distinct D has to be used for the calculation of the

corresponding rate $R(D)$ which can be calculated as a function of this Θ as well as the power spectral density of the error signal $S_{\mathrm{ee}}(\Lambda)$, which finally has to be encoded.

It is noteworthy that in contrast to the derivations from Girod for a purely translational motion model $\sigma^2_{\Delta x'}$ and $\sigma^2_{\Delta y'}$ are location-dependent for an affine motion model, since they are functions of the coordinates x and y. Consequently, $p_{\Delta X',\,\Delta Y'}(\Delta x', \Delta y')$, $P(\Lambda)$ and $S_{ee}(\Lambda)$, and finally $R(D)$ are also location-dependent.

Using the idea of generating the rate-distortion function for translative motion like explained by Girod [4] and the results from Sections II-A.1 to II-A.4, the rate-distortion function for affine motion can be evaluated, which is done in Section II-A.6.c.

6) Rate-Distortion Analysis of Affine Global Motion-Compensated Prediction: In this subsection, the minimum bit rate R (Equation (19)) for encoding the prediction error as a function of the estimation parameter variances σ^2_{e11}, σ^2_{e12}, σ^2_{e13}, σ^2_{e21}, σ^2_{e22}, σ^2_{e23} is evaluated using a fully affine motion model with 6 degrees of freedom.

For the evaluation, first in Section II-A.6.a viable affine parameter estimation errors are determined for a specific implementation using a video sequence with known frame-to-frame mappings. Based on that measurement, the displacement estimation error variances $\sigma^2_{\Delta x'}$ and $\sigma^2_{\Delta y'}$ and finally the probability density function of the displacement estimation error (according to Equation (12)) is determined. Afterwards, the gain introduced by affine motion-compensated prediction over purely translational motion-compensated prediction is analyzed. This results in the maximum gain for the special case of affine global motion-compensated prediction.

Finally in this subsection, the inherently contained non-translational affine motions in a scene ("scene affinities") of representative camera-captured aerial video sequences were measured and related to the estimation error variances.

In Section II-A.6.b the average power spectral density of real video signals is determined based on measured auto-correlation functions of different video sequences and the application of the Wiener-Khinchin theorem.

In Section II-A.6.c finally, the rate-distortion theory is applied resulting in minimum required bit rates for encoding the prediction error as a function of the affine motion parameter error variances (Fig. 3).

Without loss of generality, the computations in this subsection are carried out for *global* motion-compensated prediction, which is justified by the fact that in aerial videos from small and medium UAVs the camera-motion induced global motion is the predominant motion in each frame.

Computations for block-based motion-compensated prediction are additionally carried out in Section II-B for the simplified affine motion model.

Due to the findings of (13) and (14), the variances of the displacement estimation error $\sigma^2_{\Delta x'}$ and $\sigma^2_{\Delta y'}$ depend on the location in the frame. Consequently, also the resulting minimum achievable bit rate is location-dependent. To obtain the total bit rate for encoding one frame, the bit rate is calculated for each pel over the entire frame and subsequently summed up. Also according to (13) and (14), the variances of the displacement estimation errors $\sigma^2_{\Delta x'}$ and $\sigma^2_{\Delta y'}$ additionally depend on the variances of the error terms σ^2_{e11}, σ^2_{e12}, σ^2_{e13} for $\sigma^2_{\Delta x'}$ and on σ^2_{e21}, σ^2_{e22}, σ^2_{e23} for $\sigma^2_{\Delta y'}$, respectively.

TABLE I

MEASURED ESTIMATION ERROR VARIANCES $\sigma^2_{e_{ij}}$ IN THE ARTIFICIAL AERIAL VIDEO SEQUENCE GENERATED FROM THE *Hannover* [38] AERIAL IMAGE AS PROVIDED BY THE APPLIED AFFINE MOTION ESTIMATION IMPLEMENTATION [37].

σ^2_{e11}	σ^2_{e12}	σ^2_{e13}	σ^2_{e21}	σ^2_{e22}	σ^2_{e23}
3.27e–10	6.73e–10	3.06e−5	6.61e–10	3.19e–10	2.83e−5

a) Displacement estimation error variances, motion model error and scene "affinity": To receive viable values for the minimum bit rate R for encoding the prediction error, realistic variances σ^2_{e11}, …, σ^2_{e23} are determined (Equations (12)–(19)). Therefore, the affine estimation error variances of the affine motion estimation implementation [37] are measured. A video sequence in full High-Definition (HD) resolution of 1920×1080 pel was extracted from the aerial image *Hannover* [38] with a resolution of 10000×10000 pel. (see examples in Fig. 6 on page 7369 in the experimental section). The signal characteristic of the sequence represents realistic conditions for aerial surveillance missions. Each frame of the video sequence was generated by affine transformation (Equation (1)) of the still image *Hannover* whereas each affine parameter follows a Gaussian distribution with given means and variances, denoted as $\mathcal{N}(\text{mean};\ \text{variance})$, of:

$$\begin{aligned} &A_{11} \sim \mathcal{N}(1;\ 10^{-5})\,;\ A_{12} \sim \mathcal{N}(0;\ 10^{-5})\,;\ A_{13} \sim \mathcal{N}(15;\ 100)\,; \\ &A_{21} \sim \mathcal{N}(0;\ 10^{-5})\,;\ A_{22} \sim \mathcal{N}(1;\ 10^{-5})\,;\ A_{23} \sim \mathcal{N}(\ 0;\ \ 10)\,. \end{aligned} \tag{20}$$

A_{11}, …, A_{23} represent the random processes generating a_{11}, …, a_{23}. A Lanczos filter [39] was applied as interpolation filter. The introduced motion covers typical motion types like rotation and shearing. This sequence was used as ground truth. The variances of the estimation parameter errors of the generated video sequence are presented in Table I. These values represent the accuracy of the motion estimation implementation [37].

To analyze the overall benefit of the application of affine global motion-compensated prediction in video coding, the affine global motion parts, the "affinities", can be determined. Here, "affinity" means the inherent non-translational affine parts of the motion contained in a sequence which cannot be described in principle by a translational motion model.

If a translational motion model is used for a sequence containing a distinct affinity, the motion model error can be expressed as displacement estimation errors $\Delta x'_{\mathrm{mod}}$ and $\Delta y'_{\mathrm{mod}}$ in x- and y-direction as

$$\Delta x'_{\mathrm{mod}} = x'_{\mathrm{trans}} - x'_{\mathrm{aff}}\ ;\ \Delta y'_{\mathrm{mod}} = y'_{\mathrm{trans}} - y'_{\mathrm{aff}}. \tag{21}$$

In these two equations, x'_{trans}, y'_{trans} are the estimated displacements and x'_{aff}, y'_{aff} are the real displacements in the sequence caused by a fully affine motion inherently contained in the

scene. With a fully affine motion according to (1) (page 7362) and a purely translational motion model

$$x' = x + a_{13}\ ;\ y' = y + a_{23} \tag{22}$$

(21) yields

$$\begin{aligned}\Delta x'_{\text{mod}} &= \underbrace{(1 - a_{11})}_{e_{11,\text{mod}}} \cdot x\ \underbrace{-a_{12}}_{e_{12,\text{mod}}} \cdot y \\ &= e_{11,\text{mod}} \cdot x \quad + e_{12,\text{mod}} \cdot y \qquad (23)\\ \Delta y'_{\text{mod}} &= e_{21,\text{mod}} \cdot x \quad + e_{22,\text{mod}} \cdot y. \qquad (24)\end{aligned}$$

The parameters a_{11}, …, a_{23} in (22)–(24) are assumed to be perfectly estimated for the calculation of the motion model error, since estimation errors have already been considered separately (Table I). This means that the non-translational affine motion model errors $e_{11,\text{mod}}$, $e_{12,\text{mod}}$, $e_{21,\text{mod}}$, $e_{22,\text{mod}}$ are solely caused by motion contained in the scene which cannot be covered by a translational motion model.

The Equations (23) and (24) have the same structure as (2) and (3). Consequently, (12)–(14) also describe the motion model error if the variances of the motion model errors $\sigma^2_{e_{11,\text{mod}}}$, $\sigma^2_{e_{12,\text{mod}}}$, $\sigma^2_{e_{21,\text{mod}}}$, $\sigma^2_{e_{22,\text{mod}}}$ are inserted in (13)–(14) instead of the estimation error variances $\sigma^2_{e_{11}}$, $\sigma^2_{e_{12}}$, $\sigma^2_{e_{21}}$, $\sigma^2_{e_{22}}$. Purely translational model errors $e_{13,\text{mod}}$ and $e_{23,\text{mod}}$, or e_{13} and e_{23} in (13)–(14), respectively, are non-existent and thus set to zero.

As shown above, in case of a translational motion model, the entire "affinity" of a sequence can be considered as estimation error, since it cannot be covered by the motion model.

The affinities of four representative camera-captured aerial sequences from the TNT *Aerial Video Testset* (TAVT) data set (set 1) [13], [40] were measured. Hereby, the non-translational affine motion types (rotation, shearing, scaling) were assumed to be zero between two consecutive frames in a video sequence recorded at 30 fps and with a prevalent straight forward motion of the camera. This results in the affinities of the TAVT data set sequences as shown in Table II. From the measured results in Table II it is obvious that the variances $\sigma^2_{e_{11}}$ and $\sigma^2_{e_{22}}$ as well as $\sigma^2_{e_{12}}$ and $\sigma^2_{e_{21}}$ are pairwise similar. This can be explained by the fact that the affine motion parts are predominantly caused by a physical rotation of the camera and the skew-symmetry of a 2D rotation matrix. Justified by these findings, it is assumed that $\sigma^2_{e_{11}} = \sigma^2_{e_{22}}$ as well as $\sigma^2_{e_{12}} = \sigma^2_{e_{21}}$ and the averaged values $2.33 \cdot 10^{-7}$ and $4.63 \cdot 10^{-7}$ (see Table II), respectively, are used for further computations.

It can be observed that the variances of the model error in the range of 10^{-7} exceed the estimation error variances (approximately $5 \cdot 10^{-10}$) by several orders of magnitude. This is caused by the fact that any non-translational motion like rotation of the UAV causes a global rotation in the frame (for a camera in nadir-view) which cannot be covered by a translational motion model. Although the TAVT sequences contain prevalently straightforward motion, small rotations are also included. As a consequence also the variances of the displacement estimation errors vary by three orders of magnitude.

TABLE II

MEASURED VARIANCES $\sigma^2_{e_{ij}}$ OF NON-TRANSLATIONAL AFFINE TRANSFORMATION PARAMETERS ("AFFINITY") OF AERIAL VIDEOS FROM THE TNT *Aerial Video Testset* (TAVT) DATA SET (SET 1) [13], [40]. THE SEQUENCE (SEQ.) NAMES REFER TO THE FLIGHT ALTITUDES THEY WERE RECORDED AT.

Seq. name	$\sigma^2_{e_{11}}$	$\sigma^2_{e_{12}}$	$\sigma^2_{e_{21}}$	$\sigma^2_{e_{22}}$	Mean $(\sigma^2_{e_{11}}, \sigma^2_{e_{22}})$	Mean $(\sigma^2_{e_{12}}, \sigma^2_{e_{21}})$
350 m seq.	2.03e−7	6.03e−7	6.59e−7	2.24e−7	2.13e−7	6.31e−7
500 m seq.	1.94e−7	5.09e−7	3.63e−7	1.94e−7	1.94e−7	4.35e−7
1000 m seq.	1.74e−7	4.05e−7	4.13e−7	2.12e−7	1.93e−7	4.09e−7
1500 m seq.	3.19e−7	3.80e−7	3.69e−7	3.46e−7	3.33e−7	3.75e−7
Mean	2.23e−7	4.74e−7	4.51e−7	2.44e−7	2.33e−7	4.63e−7

b) Power spectral density of the video signal: For the calculation of the power spectral density S_{ss} of the video signal, the exponential drop rates α_x and α_y of the autocorrelation function are required (Equation (16)). Thus, the mean correlations [36] of horizontally and vertically adjacent pels of several video sequences from the *Joint Collaborative Team on Video Coding* (JCT-VC) test set [41] were calculated. For the standard-definition (SD) sequences (720 × 576) *OldTownCross*, *CrowdRun*, *ParkJoy*, *DucksTakeOff*, and *InToTrees* mean horizontal and vertical correlations of $\rho_{ss,x} = 0.9425$ and $\rho_{ss,y} = 0.9266$, respectively, were measured. For the HD sequences (1920 × 1080) *BasketballDrive*, *BQTerrace*, *Cactus*, *Kimono*, and *ParkScene* the averaged horizontal and vertical correlations amount to $\rho_{ss,x} = 0.9744$ and $\rho_{ss,y} = 0.9677$, respectively. It can be observed that the correlations between adjacent pels are larger for higher resolution sequences (HD) compared to lower resolution sequences as those used by Girod. Since the video characteristics have not fundamentally changed and comparable focal lengths were used for capturing, much more pels represent one object in a HD sequence than in a low resolution sequence (e. g. QCIF, CIF, or SD [3]) and consequently, the correlations between pels have to be higher for HD sequences. The Fourier transform of the autocorrelation function now is the power spectral density of the signal S_{ss} according to the Wiener-Khinchin theorem as explained above.

c) Application of the rate-distortion theory: The evaluation of the rate-distortion theory (Equations (18) and (19)) yields the minimum required bit rate R for a distortion D. The location-dependent bit rate is visualized in Fig. 2 for a HD resolution frame with non-translational affine estimation error variances of $\sigma^2_{e_{11}} = \sigma^2_{e_{12}} = \sigma^2_{e_{21}} = \sigma^2_{e_{22}} = 5 \cdot 10^{-10}$ (cf. Table I), translational estimation error variances $\sigma^2_{e_{13}} = \sigma^2_{e_{23}} = 0$, and Θ selected to yield a signal-to-noise ratio (SNR) of 30 dB. In Fig. 3 the bit rate is plotted versus the translational variances on one axis ($\sigma^2_{e_{13}}$, $\sigma^2_{e_{23}}$) and the non-translational affine variances ($\sigma^2_{e_{11}}$, $\sigma^2_{e_{12}}$, $\sigma^2_{e_{21}}$, $\sigma^2_{e_{22}}$) on the other axis. For visualization both translational and all non-translational affine error variances are assumed to be equal. Isolines are marked by data tips in the 3D plot in Fig. 3 for a translational half-pel resolution (data tip for "transl. var.: 0.0208") as well as quarter-pel resolution (data tips with "transl. var.: 0.0052") and

[3] QCIF: quarter common intermediate format (resolution of 176×144); CIF: common intermediate format (resolution of 352×188); SD: standard-definition (resolution of 720×576 for the phase alternating line system (PAL)).

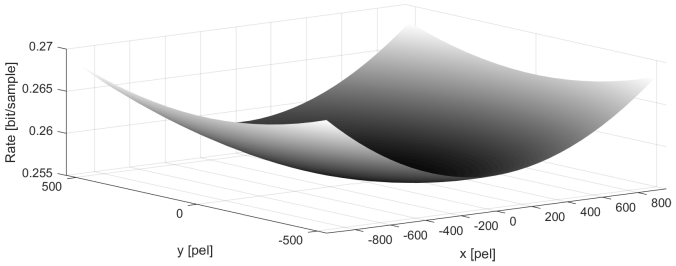


Fig. 2. Location-dependent bit rate for a HD frame and maximum accurate non-translational affine estimation ($\sigma^2_{e11}=\sigma^2_{e12}=\sigma^2_{e21}=\sigma^2_{e22}=5\cdot10^{-10}$) and translational quarter-pel resolution.

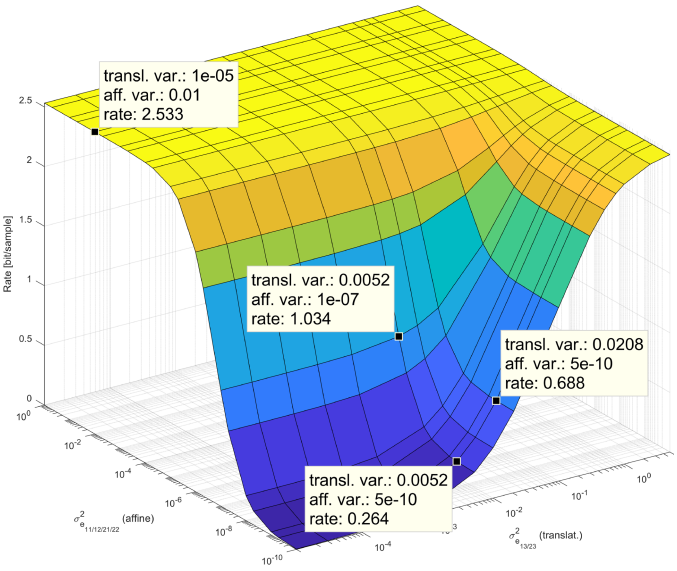


Fig. 3. Minimum required bit rate versus variances $\sigma^2_{e_{ij}}$ for a distortion of SNR = 30 dB assuming $\sigma^2_{e11}=\sigma^2_{e12}=\sigma^2_{e21}=\sigma^2_{e22}$ (called $\sigma^2_{e_{\text{aff.}}}$ in (b)) and $\sigma^2_{e13}=\sigma^2_{e23}$ for HD resolution.

non-translational affine estimation error variances of $\sigma^2_{e11}=\sigma^2_{e12}=\sigma^2_{e21}=\sigma^2_{e22}=5\cdot10^{-10}$ (cf. Table I).

7) Conclusions for the Fully Affine Motion Model for Global Motion Compensation: From the results it can be inferred:

- The variances of the estimation errors of the non-translational affine parameters (σ^2_{e11}, σ^2_{e12}, σ^2_{e21}, σ^2_{e22}) have to be magnitudes smaller than the variances of the translational parameters (σ^2_{e13}, σ^2_{e23}) to yield reasonably small bit rates. For a potential quantization of the non-translational affine parameters for encoding purposes this fact should be taken into account. The error variances as well as the bit rates are location-dependent, which becomes important for non (purely) translational motion like rotation.
- The isoline with all non-translational affine error variances equal to zero (not printed in the logarithmic plot in Fig. 3) describes the bit rate for encoding the prediction error for a translational motion model (which is identical to the results from Girod [4] for same correlations). Non-translational affine variances unequal to zero obviously can only occur if an affine model is employed. In such a case, affine motions contained in a scene can be matched much better than with a purely translational motion model, i.e. the operating point moves towards the dark blue plateau in Fig. 3. Using an affine motion model is especially beneficial in the case that high amounts of non-translational motions are contained in a scene.
- For a sequence with a specific degree of non-translational affine motion ("affinity"), which cannot be described by a translational motion model, the minimum bit rate is limited along the (non-translational) affine-variances-axis (directing from the origin leftwards in Fig. 3). As an example, a HD sequence with an "affinity" of 10^{-7} is assumed (Table II). The additional estimation error is negligible in this example since it is three orders of magnitude smaller (see Table I) and consequently also the contribution of the estimation error to the bit rate is negligible. For the example above the minimum bit rate for encoding the prediction error using a purely translational motion estimation with the small estimation error variances of $\sigma^2_{e13}=\sigma^2_{e23}=0.0052$ is 1.034 bit/sample (central data tip in Fig. 3). In contrast to that the minimum bit rate is only 0.264 bit/sample for an accurate *affine* motion estimation with estimation error variances of $\sigma^2_{e11}=\sigma^2_{e12}=\sigma^2_{e21}=\sigma^2_{e22}=5\cdot10^{-10}$ and the same translational accuracy of 1/4-pel resolution (lower data tip in Fig. 3).
- From the example given in the third bullet point, it can be generalized that the minimum required bit rate is reached, if the motion model covers the real motion contained in the scene, *and* if the affine estimation is highly accurate. The feasibility of this requirement is shown in this work.
- As it is obvious from (12)–(14), $\sigma^2_{\Delta x'}$ and $\sigma^2_{\Delta y'}$ increase for large image dimensions. For block-based motion compensation, the "frame dimensions" are equal to the block dimensions. A block-based affine motion-compensated prediction is analyzed in the following subsection.

B. Efficiency Analysis of Simplified Affine Motion Compensation

An efficiency analysis of a fully affine motion model has been presented in the previous subsection. In contrast to that, a simplified affine motion model with only 4 degrees of freedom is assumed here. Although "simplified" in the name suggests that also the theoretical analysis is simplified, additional dependencies between the parameters of the model have to be considered. However, the basic structure of the derivation remains the same and only the modeling of the probability density function ${}^{\text{simp}}p_{\Delta X'_{\text{s}},\Delta Y'_{\text{s}}}\left(\Delta x'_{\text{s}},\Delta y'_{\text{s}}|x,y\right)$ is different.

1) Derivation of the Probability Density Function of the Displacement Estimation Error for a Simplified Affine Model: A simplified affine model with four parameters like proposed by Li *et al.* [14] is assumed.

With the rotation angle θ, the scaling factor s_{s} in both, horizontal and vertical direction, and the translational parameters c and f (which correspond to the parameters a_{13} and a_{23} in the fully affine model in Section II-A.1), the relationship

between the coordinates x and y before and x'_s and y'_s after the transformation is described in [14] as

$$\begin{aligned} x'_s &= \quad s_s \cos\theta \cdot x + s_s \sin\theta \cdot y + c \ ; \\ y'_s &= -s_s \sin\theta \cdot x + s_s \cos\theta \cdot y + f. \end{aligned} \tag{25}$$

Replacing

$$(s_s \cos\theta) \text{ by } (1+a) \quad \text{and} \quad (s_s \sin\theta) \text{ by } b \tag{26}$$

respectively, (25) can be expressed as

$$\begin{aligned} x'_s &= (a+1) \cdot x + \quad b \cdot y + c \ ; \\ y'_s &= -b \cdot x + (a+1) \cdot y + f. \end{aligned} \tag{27}$$

Each parameter a, b, c, f is assumed to be perturbed by an independent error term e_i, with $i = \{a, b, c, f\}$, caused by inaccurate parameter estimation. The perturbed coordinates $\hat{x}_s$, $\hat{y}_s$ lead to displacement estimation errors in horizontal and vertical direction of $\Delta x'_s$ and $\Delta y'_s$ (in pel)

$$\begin{aligned} \Delta x'_s &= \hat{x}'_s - x'_s = \quad e_a \cdot x + e_b \cdot y + e_c \ ; \\ \Delta y'_s &= \hat{y}'_s - y'_s = -e_b \cdot x + e_a \cdot y + e_f. \end{aligned} \tag{28}$$

Assuming each error term e_i to be zero-mean Gaussian distributed leads to the probability density functions (pdfs)

$$p(e_i) = \frac{1}{\sqrt{2\pi\,\sigma_{e_i}^2}} \cdot \exp\left(-\frac{e_i^2}{2\sigma_{e_i}^2}\right) \text{ with } i = \{a, b, c, f\}. \tag{29}$$

For statistically independent variables, the joint pdf $p_{E_a,E_b,E_c,E_f}(e_a, e_b, e_c, e_f)$ for the random processes E_a, E_b, E_c, E_f and the observations e_a, e_b, e_c, e_f is:

$$p_{E_a,E_b,E_c,E_f}(e_a, e_b, e_c, e_f) = p(e_a) \cdot p(e_b) \cdot p(e_c) \cdot p(e_f). \tag{30}$$

In order to convert the pdf $p_{E_a,E_b,E_c,E_f}(e_a, e_b, e_c, e_f)$ to the desired pdf ${}^{\text{simp}}p_{\Delta X'_s,\Delta Y'_s}(\Delta x'_s, \Delta y'_s|x, y)$, the transformation theorem for pdfs can be used again (Equation (6)). Here, $\Delta X'_s$, $\Delta Y'_s$ are the random processes generating the displacement estimation errors $\Delta x'_s$, $\Delta y'_s$ (in pel) caused by affine parameter estimation errors. With (28) this yields

$$\begin{aligned} &{}^{\text{simp}}p_{\Delta X'_s,\Delta Y'_s}\left(\Delta x'_s, \Delta y'_s|x, y\right) \\ &= \int_{\mathbf{R}^4} p_{E_a,\ldots,E_f}(e_a, \ldots, e_f) \\ &\quad \cdot \delta\big(\Delta x'_s - (e_a x + e_b y + e_c)\big) \\ &\quad \cdot \delta\big(\Delta y'_s - (-e_b x + e_a y + e_f)\big)\ \mathrm{d}e_a \mathrm{d}e_b \mathrm{d}e_c \mathrm{d}e_f \end{aligned} \tag{31}$$

with a dependency on the location coordinates x, y in the current frame. Using the properties of the delta function results in

$$\begin{aligned} &{}^{\text{simp}}p_{\Delta X'_s,\Delta Y'_s}\left(\Delta x'_s, \Delta y'_s|x, y\right) \\ &= \int_{\mathbf{R}^2} p_{E_a,E_b,E_c,E_f}(e_a, e_b, \Delta x'_s - e_a x - e_b y, \\ &\quad \Delta y'_s + e_b x - e_a y)\ \mathrm{d}e_a \mathrm{d}e_b. \end{aligned} \tag{32}$$

Considering (32) and (29) results in:

$$\begin{aligned} &{}^{\text{simp}}p_{\Delta X'_s,\Delta Y'_s}\left(\Delta x'_s, \Delta y'_s|x, y\right) = \frac{1}{(2\pi)^2 \sigma_{e_a}\sigma_{e_b}\sigma_{e_c}\sigma_{e_f}} \\ &\cdot \int_{\mathbf{R}^2} \exp\Bigg(-\frac{e_a^2}{2\sigma_{e_a}^2} - \frac{e_b^2}{2\sigma_{e_b}^2} - \frac{\left(\Delta x'_s - e_a x - e_b y\right)^2}{2\sigma_{e_c}^2} \\ &\quad - \frac{\left(\Delta y'_s + e_b x - e_a y\right)^2}{2\sigma_{e_f}^2}\Bigg)\ \mathrm{d}e_a \mathrm{d}e_b. \end{aligned} \tag{33}$$

After the two integrations

$$\begin{aligned} &{}^{\text{simp}}p_{\Delta X'_s,\Delta Y'_s}\left(\Delta x'_s, \Delta y'_s|x, y\right) \\ &\quad = \frac{1}{2\pi\sqrt{N}} \cdot \exp\left(\frac{M}{2N}\right) \end{aligned} \tag{34}$$

$$\begin{aligned} \text{with } N = &\left(\left(x^2+y^2\right)^2 \sigma_{e_b}^2 + y^2\sigma_{e_c}^2 + x^2\sigma_{e_f}^2\right)\sigma_{e_a}^2 \\ &+ \left(x^2\sigma_{e_c}^2 + y^2\sigma_{e_f}^2\right)\sigma_{e_b}^2 + \sigma_{e_c}^2\sigma_{e_f}^2 \end{aligned} \tag{35}$$

$$\begin{aligned} \text{and } M = &-\left(x\Delta y'_s - y\Delta x'_s\right)^2 \sigma_{e_a}^2 - \left(x\Delta x'_s + y\Delta y'_s\right)^2 \sigma_{e_b}^2 \\ &- \Delta x'^2\sigma_{e_f}^2 - \Delta y'^2\sigma_{e_c}^2 \end{aligned} \tag{36}$$

is obtained.

Transforming (34) into the form of a common bivariate zero-mean normal distribution with ρ being the correlation coefficient between $\Delta X'$ and $\Delta Y'$ leads to the desired final pdf of the displacement estimation error:

$$\begin{aligned} &{}^{\text{simp}}p_{\Delta X'_s,\Delta Y'_s}\left(\Delta x'_s, \Delta y'_s|x, y\right) \\ &\quad = \frac{1}{2\pi\ \sigma_{\Delta x'_s}\ \sigma_{\Delta y'_s}\sqrt{1-\rho^2}} \\ &\qquad \cdot \exp\left(-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{\Delta x'^2_s}{\sigma^2_{\Delta x'_s}} + \frac{\Delta y'^2_s}{\sigma^2_{\Delta y'_s}} - \frac{2\rho \cdot \Delta x'_s \cdot \Delta y'_s}{\sigma_{\Delta x'_s} \cdot \sigma_{\Delta y'_s}}\right]\right) \end{aligned} \tag{37}$$

$$\begin{aligned} &\text{with } \sigma^2_{\Delta x'_s} \\ &\quad = N \cdot \Big(\left(\sigma_{e_a}^2 y^2 + \sigma_{e_b}^2 x^2 + \sigma_{e_f}^2\right) \cdot \left(1-\rho^2\right)\Big)^{-1}, \end{aligned} \tag{38}$$

$$\sigma^2_{\Delta y'_s} = N \cdot \Big(\left(\sigma_{e_a}^2 x^2 + \sigma_{e_b}^2 y^2 + \sigma_{e_c}^2\right) \cdot \left(1-\rho^2\right)\Big)^{-1} \tag{39}$$

$$\rho = \frac{\left(\sigma_{e_a}^2 xy - \sigma_{e_b}^2 xy\right)}{\sqrt{\sigma_{e_a}^2 y^2 + \sigma_{e_b}^2 x^2 + \sigma_{e_f}^2}\ \sqrt{\sigma_{e_a}^2 x^2 + \sigma_{e_b}^2 y^2 + \sigma_{e_c}^2}}. \tag{40}$$

Obviously, the variances $\sigma^2_{\Delta x'_s}$ and $\sigma^2_{\Delta y'_s}$ depend on the locations x, y in the frame similarly to the fully affine model (Section II-A.1). Further on, ${}^{\text{simp}}p_{\Delta X'_s,\Delta Y'_s}\left(\Delta x'_s, \Delta y'_s|x, y\right)$ is abbreviated as ${}^{\text{simp}}p_{\Delta X'_s,\Delta Y'_s}\left(\Delta x'_s, \Delta y'_s\right)$ for simplicity. Moreover, in contrast to the fully affine case in Section II-A, the variances of the random processes $\Delta X'_s$ and $\Delta Y'_s$ both depend on the variances of *all* estimated parameters and thus $\Delta X'_s$ and $\Delta Y'_s$ are interdependent.

For equal variances $\sigma_{e_a}^2 = \sigma_{e_b}^2$, the correlation coefficient ρ becomes zero, since the influence on $\Delta x'_s$ and $\Delta y'_s$ is pairwise similar. Thus $\Delta x'$ and $\Delta y'$ can be considered as uncorrelated and the pdfof the displacement estimation error

TABLE III

MEASURED VARIANCES $\sigma^2_{e_a}$, $\sigma^2_{e_b}$ [14] OF SIMPLIFIED AFFINE TRANSFORMATION PARAMETERS OF AERIAL VIDEOS FROM THE TAVT DATA SET (SET 1) [13], [40].

	$\sigma^2_{e_a}$	$\sigma^2_{e_b}$
350 m sequence	$1.92 \cdot 10^{-7}$	$6.23 \cdot 10^{-7}$
500 m sequence	$1.86 \cdot 10^{-7}$	$3.74 \cdot 10^{-7}$
1000 m sequence	$1.79 \cdot 10^{-7}$	$4.06 \cdot 10^{-7}$
1500 m sequence	$3.21 \cdot 10^{-7}$	$3.67 \cdot 10^{-7}$
Mean	$2.20 \cdot 10^{-7}$	$4.43 \cdot 10^{-7}$

of the simplified affine model becomes the solution for the fully affine case.

2) Rate-Distortion Analysis of the Simplified Affine Model: To derive the minimum required bit rate for encoding the prediction error using motion-compensated prediction in video coding using the simplified affine model from Section II-B.1, the derivations from Section II-A.4 and Section II-A.5 are employed again. With the Fourier transform ${}^{\text{simp}}P(\Lambda)$ of ${}^{\text{simp}}p_{\Delta X'_s, \Delta Y'_s}\left(\Delta x'_s, \Delta y'_s\right)$ from the last subsection and Equation (17), the power spectral density of the prediction error ${}^{\text{simp}}S_{ee}(\Lambda)$ for the simplified affine model is derived. Hereby, the same power spectral density of the signal S_{ss} is assumed as derived in Section II-A.3. Evaluating the rate-distortion theory by exploiting (18) and (19) yields the distortion D and the minimum required bit rate $R(D)$, which correspond to ${}^{\text{simp}}D$ and ${}^{\text{simp}}R\left({}^{\text{simp}}D\right)$, respectively, for encoding the prediction error by using a simplified affine model as defined in (27). The rate-distortion theory for the simplified affine motion-compensated prediction is evaluated in accordance with the procedure described in Section II-A.6, where the analysis was carried out for the fully affine model. For evaluation, the same autocorrelation function (Equation (15)) of the signal as determined in Section II-A.6.b was assumed. As discussed above, the only difference in the evaluation is that the Fourier transform of the pdf of the displacement estimation error from the simplified affine model ${}^{\text{simp}}P(\Lambda)$ is inserted in (17) instead of the Fourier transform of the pdfof the displacement estimation error from the fully affine model $P(\Lambda)$.

Evaluation of the rate-distortion theory for a distortion of SNR = 30 dB results in minimum required bit rates for different variances $\sigma^2_{e_i}$ of Gaussian displacement estimation error pdfs of the simplified affine transformation parameters as shown in Fig. 4. For the simulations, the affine parameters were assumed to be in the fixed ratio $\sigma^2_{e_b} = 2\sigma^2_{e_a}$ as measured (see Table II) and both translational parameters to be equal ($\sigma^2_{e_c} = \sigma^2_{e_f}$).

The affinities of the aerial video sequences from the TAVT data set (set 1) were measured similarly to the measures presented in Section II-A.6.a. The results are given in Table III and support the ratio. It is obvious that the results for the simplified affine model are almost the same as for the fully affine model (Table II) since barely no motions are contained in the sequences which cannot be covered by the simplified model. Moreover the smaller number of parameters of the simplified model may be estimated more accurately.

The minimum bit rates as a function of the simplified non-translational affine (axis from center to left) and translational

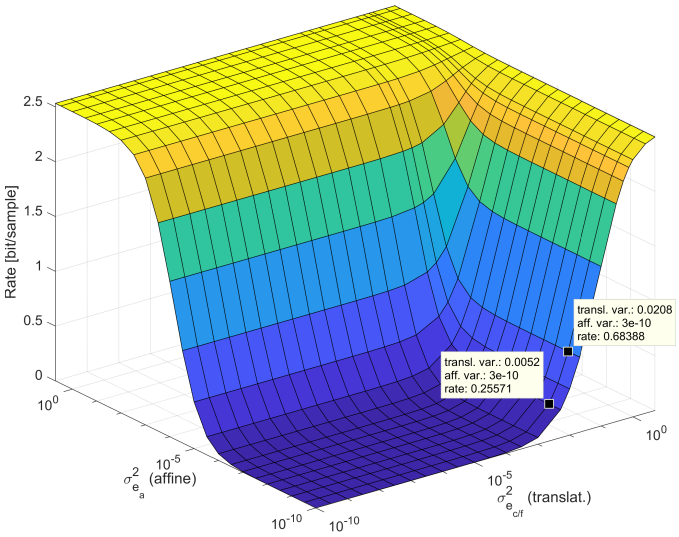
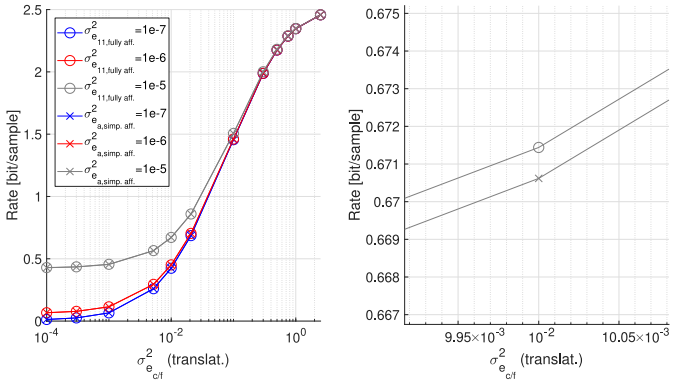


Fig. 4. Minimum required bit rate versus variances $\sigma^2_{e_i}$, $i = a, b, c, f$ of Gaussian displacement estimation error pdfs for SNR = 30 dB assuming $\sigma^2_{e_b} = 2\sigma^2_{e_a}$ and $\sigma^2_{e_c} = \sigma^2_{e_f}$ (block size 64×64 pel, transform center in the middle of the block) [30].

variances (axis from center to right) are presented in Fig. 4 for a block size of 64×64 pel.

It is noteworthy that the operating point ($\sigma^2_{e_b} = 2\sigma^2_{e_a}$, $\sigma^2_{e_c} = \sigma^2_{e_f}$) reaches higher bit rates if the motion contained in the sequence cannot be represented by the motion model. This is the case when a purely translational motion model is used to estimate a sequence containing distinct (non purely translational) affine motion—albeit the resulting bit rate difference decreases for smaller block (or frame) sizes. For the example of a block size of 64×64 pel, the minimum required bit rate for an accurate simplified affine estimation of $\sigma^2_{e_a} = 3 \cdot 10^{-10}$, $\sigma^2_{e_b} = 6 \cdot 10^{-10}$ and $\sigma^2_{e_c} = \sigma^2_{e_f} = 3 \cdot 10^{-5}$ amounts to 0.0020 bit/sample. For a translational quarter-pel resolution and equal non-translational affine variances, the bit rate increases to 0.2557 bit/sample (lower data tip in Fig. 4). On the contrary, for a purely translational motion model with the same translational quarter-pel resolution, the non-translational affine part of the motion contained in the scene cannot be covered at all, leading to high variances $\sigma^2_{e_a} = 2.2 \cdot 10^{-7}$, $\sigma^2_{e_b} = 4.4 \cdot 10^{-7}$ and consequently higher bit rates of 0.2645 bit/sample for block sizes of 64×64 pel or 1.5589 bit/sample for global motion compensation (both not shown). For the example of translational quarter-pel resolution (which is equal to translational estimation error variances of 0.0052 as already stated), a block size of 64×64 pel and a ratio of the non-translational simplified affine estimation errors of $\sigma^2_{e_b} = 2\sigma^2_{e_a}$, the rate-distortion optimized bit rate for intra encoding of the HD resolution video signal itself amounts to a bit rate of 1.9918 bit/sample. Considering bit rates and corresponding estimation error variances for encoding the prediction error (Fig. 4), it can be concluded that simplified affine motion-compensated prediction achieves improvements for non-translational affine variances of about $\sigma^2_{e_a} = 3 \cdot 10^{-4}$ or smaller.

In Fig. 5 the bit rates are compared for a fully affine model (six degrees of freedom) (circles) and the simplified, four-



(a) Simplified vs. fully affine (b) Magnification

Fig. 5. Minimum required bit rate and achievable gains of the simplified affine vs. the fully affine motion model for a block size of 64×64 pel (SNR = 30 dB, $\sigma^2_{e_b} = 2\sigma^2_{e_a}$, $\sigma^2_{e_c} = \sigma^2_{e_f}$), magnification in (b) [30].

(a) Full image, 10000×10000 pel (b) Cropped full HD image

Fig. 6. Test image *Hannover* [38].

parameter model (crosses) for 64×64 pel blocks as used as maximum block size in the current video coding standard HEVC. The plots show that the simplified model requires a smaller amount of bits for encoding the prediction error compared to a fully affine model for equal error variances. This can be explained since in (40) it became obvious that X'_s and Y'_s are correlated for the simplified affine model. On the other hand, the model error may increase for sequences containing motions which cannot be covered by a simplified but by a fully affine model, which occurs, e. g. for shearing. However, the difference between the simplified and the fully affine model is negligible in terms of bit rate saving. Motions which cannot be covered by the simplified affine motion model rarely occur in surveillance video sequences—and presumably also in general videos only to a minor extent. Thus, from a coding point of view it is beneficial to encode as few parameters as possible and consequently, the use of the simplified affine model for encoding purposes is reasonable.

III. Experiments

This section is divided into two parts: in Section III-A the rate-distortion theory for affine motion-compensated prediction from Section II is investigated. The unquantized prediction error is quantized so that a predefined distortion, e. g. of 30 dB signal-to-noise ratio (SNR), is introduced between the original signal and the quantization error. In this manner, the prediction error caused by inaccurate motion estimation as modeled in Section II is simulated and the prediction error bit rate model is validated.

Operational rate-distortion diagrams for common video test sequences containing distinct non-translational affine motions are presented and compared to those of real-world aerial and non-aerial video sequences in Section III-B.

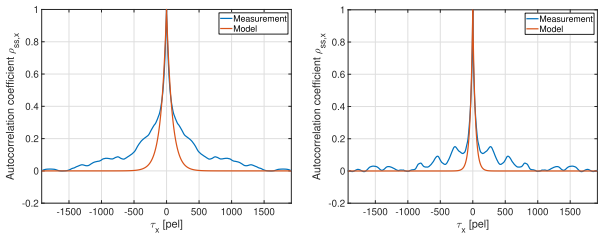
A. Affine Motion Compensation in Video Coding

In this subsection the model for calculating the minimum required bit rate for encoding the prediction error using affine motion-compensated prediction from Section II is verified. Moreover, video sequences with and without distinct affine—non purely translational—motion are compared.

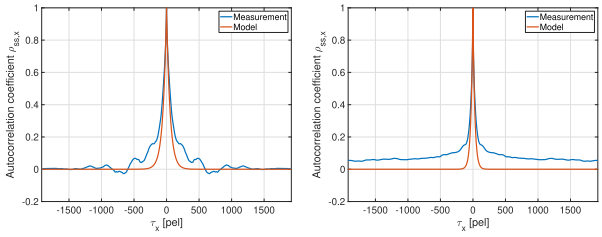
In Section II the rate-distortion function for affine motion-compensated prediction was derived and the bit rate for encoding the prediction error was calculated as a function of the motion estimation accuracy. The latter was characterized by the variances of errors of the affine mapping parameters. For visualization of the results in Fig. 3 in a 3D plot, the variances of the errors of the translational affine parameters $\sigma^2_{e_{13}}$, $\sigma^2_{e_{23}}$ and the non-translational affine parameters $\sigma^2_{e_{11}}$, $\sigma^2_{e_{12}}$, $\sigma^2_{e_{21}}$, $\sigma^2_{e_{22}}$, respectively, were assumed to be equal. Since the Gaussian distribution has the highest entropy among all distributions with given mean and variance—and Gaussian distributions have been assumed for the distributions of the motion estimation errors—the resulting model bit rate is the supremum of the minimum required bit rate for encoding the prediction error. In other words, for any non-Gaussian distribution, the rate-distortion optimized minimum required bit rate for encoding the prediction error is expected to be smaller than predicted by the model.

To validate the model introduced in Section II-A, the 10000×10000 pel aerial image of Hannover (Fig. 6) [38] was used. The image provides a similar signal characteristic as the other HD test sequences in terms of its autocorrelation coefficients. In Fig. 7 the autocorrelation coefficient of *Hannover* is compared to those of a HD resolution JCT-VC test sequence [10], a test sequence which contains high amounts of non-translational affine motions proposed by Li [14], and an aerial test sequence from the TAVT data set [13], [40]. The plots show that the autocorrelation coefficients almost perfectly match the model assumed in Section II-A.3 for small and medium pel shifts of $|\tau_x| \leq 50$.

A virtual camera has been used to extract several full HD resolution (1920×1080 pel) frames from the large aerial image of Hannover which were concatenated to a video sequence. The frame-to-frame motions comply with an affine motion model (Section II-A.1). To generate the affine motion for the virtual camera path, (pseudo-) random numbers were drawn from a Gaussian distribution with given means and variances. The means of the affine parameters are selected such that the mean frame-to-frame motion is zero, i. e. the mean value is



(a) Sequence *BasketballDrive* (b) Sequence *ShieldsPart*



(c) Sequence *350 m sequence* (d) Test image *Hannover*

Fig. 7. Measured autocorrelation coefficients $\rho_{ss,x}$ in horizontal direction. of the natural video test sequences *BasketballDrive* (Fig. 11a) [10] and *ShieldsPart* (Figs. 10a and 10b) [14] as well as of the aerial test sequence *350m sequence* (Fig. 12a) from the TAVT data set [13], [40] (averaged over 50 frames each) and of the test image *Hannover* (10000×10000 pel) [38] (cropped to same width, Fig. 6b). For the model evaluation in the plots (red lines), the correlation coefficient of each specific test sequence or image, respectively, was used. It can be seen that the exponentially decaying model perfectly fits for small to medium shifts up to ± 50 pel and that the measured correlations are small (≤ 0.4) for larger shifts.

1 for the parameters a_{11} and a_{22}, and zero for the remaining parameters a_{12}, a_{13}, a_{21}, a_{23}. The resulting video sequence signal s now has well-known frame-to-frame mappings. Thus, the sequence can be used to measure the bit rates needed for encoding the prediction error. Assuming the most trivial motion estimation system which always predicts "no motion", the artificially introduced motion becomes exactly the prediction error e, which can be calculated just as the difference between two consecutive frames in the video sequence.

The setup for the measurement is shown in Fig. 8. The unquantized prediction error e is decorrelated using a differential pulse-code modulation (DPCM), where only the correlations between horizontally ($s_{k_{x-1,y}}$) and vertically ($s_{k_{x,y-1}}$) neighboring pels are exploited for the prediction of the current pel $\hat{s}_{k_{x,y}}$ at position (x, y), whereas $\hat{s}_{k_{x,y}} = s_{k_{x,y}} - 0.5\ s_{k_{x-1,y}} - 0.5\ s_{k_{x,y-1}}$. An uniform quantization is applied afterwards so that the signal-to-noise ratio between the video signal s and the quantization error $e_\mathrm{q} = e' - e$ is equal to a predefined value, e. g. of 30 dB as assumed for the model in Section II-A.6.c. The bit rate is calculated as the entropy of a memoryless source, which corresponds to the bit rate needed for encoding the quantized prediction error, assuming perfectly decorrelated symbols after the DPCM.

The results are shown in Fig. 9 for a SNR of 30 dB (like assumed in the entire Section II) and using 30 frames with different motions for each data point. It is obvious that the measurement qualitatively perfectly matches the theory, but that the measured bit rates are smaller than those of the model (Fig. 3). For instance, the measured maximum bit rate is 2.507 bit/sample, for 10^{-5} for the translational variances

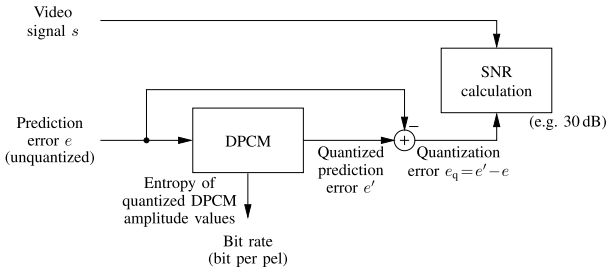


Fig. 8. Setup for measuring the bit rate for encoding the prediction error. The quantization is adjusted so that a predefined distortion, e. g. 30 dB SNR, is met and the corresponding entropy of the quantized DPCM amplitude values is determined.

$\sigma^2_{e_{13}}$, $\sigma^2_{e_{23}}$ and 10^{-2} for the non-translational affine variances $\sigma^2_{e_{11}}$, $\sigma^2_{e_{12}}$, $\sigma^2_{e_{21}}$, $\sigma^2_{e_{22}}$ (upper left data tip in Fig. 9), instead of 2.53 bit/sample as predicted by the model (upper left data tip in Fig. 3). Accordingly for the translational variances of 0.0052 and the non-translational affine variances of 10^{-7}, the measured bit rate of 0.382 bit/sample (central data tip in Fig. 9) is lower than the model bit rate of 1.034 bit/sample (central data tip in Fig. 3). For very small variances below 0.02 (translational) and 10^{-8} (non-translational), the measured bit rates faster decrease to zero than predicted by the model (dark blue plateau in Fig. 9). These differences may mainly be caused by the low-pass filtering character of the Lanczos interpolation filter used during the generation of the test sequence *Hannover*. By low-pass filtering, higher frequencies are flattened or entirely removed from the signal, which finally leads to smaller bit rates needed for encoding the prediction error in the measurement compared to the model. The pronounced lower plateau mainly occurs since for very small affine motions introduced during the generation of the test sequence, the affine distorted image perfectly matches the original after interpolation filtering. Thus, the prediction error image as introduced above is nil and consequently no bit rate is needed for encoding it.

Since the model bit rates represent the supremum of the minimum required bit rate for encoding the prediction error, the measurements empirically prove the correctness of the model.

To reveal the operating range of the model for real-world sequences, the test sequence *ShieldsPart* (1920×1080, 50 fps, 8 bit/sample, chroma subsampling 4:2:0, 100 frames) (Fig. 10) [14] was exemplary used. It was encoded using JEM 7.1 (*Apache Subversion* (SVN) revision 603) [16] with the *random access* (RA) profile and the *low-delay p* (LDP) profile [24]. The sequence has been proposed to demonstrate the efficiency of affine motion-compensated prediction by Li *et al.* [14] since it contains distinct (non-translational) affine motion. Taking the average luminance value of 49.5 for *ShieldsPart* into account, a SNR of 30 dB corresponds to a PSNR of 44.2 dB on an 8-bit scale. The averaged bit rate over both profiles, only using non-intra coded frames for encoding the sequence at the given PSNR for the luminance component is 68599 kbit/s, which corresponds to a mean bit rate of 0.66 bit/sample. This bit rate also includes signaling, which is neither covered by the model

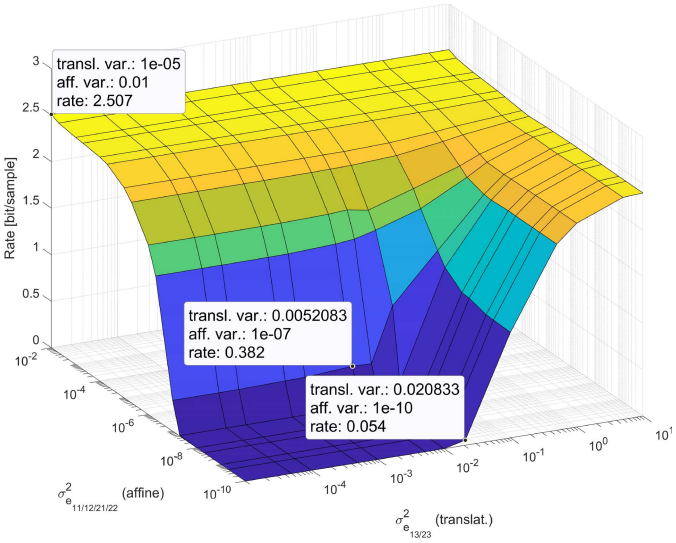


Fig. 9. Bit rate for encoding the prediction error as a function of the motion estimation error variances for a frame in full HD resolution (1920×1080) using DPCM for signal decorrelation and uniform quantization.

(a) *ShieldsPart*, frame 1 (b) *ShieldsPart*, frame 100 (last frame)

Fig. 10. Test sequence *ShieldsPart* (50 fps) containing distinct (non-translational) affine motion [14].

nor considered in the measurement. Extrapolating the findings for AVC from Klomp [31] to HEVC or, more specific, to the HEVC *Test Model*- (HM-) based JEM software, the signaling data may account for less than 10 percent in the case of fine quantization (*quantization parameter* (QP) 15 for RA and QP 16 for LDP).

As can be seen in Fig. 9, the operating point for the sequence is located in the middle of the mid-blue area above the marked point at "transl. var.: 0.0052; aff. var.: 1e-7" (central data tip in Fig. 9) for the sequence. Hereby, translational quarter-pel resolution like used in the JEM software [42] is assumed, which corresponds to the isoline at the translational variance 0.0052. Using the derivations of Section II-A, the model bit rate for the sequence is approximately 2.2 $\frac{\text{bit}}{\text{sample}}$ (not shown) for a block size of 128×128 pel as used in JEM with translational quarter-pel resolution and non-translational affine motion vector accuracies of $\frac{1}{16}$ pel, which corresponds to the internal luma resolution of JEM. Compared to the modeled bit rate, the measured bit rate is approximately half as high, which can be explained as follows: first, the assumption of a stationary signal was made in the model, which may not entirely be true for natural videos. Second, in the example calculations the translational and the non-translational parameter error variances were each assumed to be identical, i. e. $\sigma_{e_{13}}^2 = \sigma_{e_{23}}^2$ and $\sigma_{e_{11}}^2 = \sigma_{e_{12}}^2 = \sigma_{e_{21}}^2 = \sigma_{e_{22}}^2$, which may not be fulfilled in case that the error variances are not predominantly caused by artificial quantization of the (simplified) affine parameters. Third, the autocorrelation function of the signal was assumed to be exponentially decreasing, which is a good approximation of a video signal although it is not entirely reflecting reality. Especially for larger shifts it was demonstrated in Fig. 7 that the exponentially decreasing autocorrelation model tends to overestimate the high frequency components contained in the signal, which increases the bit rate in the model. Moreover, in the model calculations averaged correlations between adjacent pels were assumed to generalize the model. However, for specific sequences these correlations may highly vary, leading to different modeled bit rates (cf. Section II-A.6.b). Finally, the displacement estimation error pdfwas derived to be Gaussian distributed, induced by the assumed Gaussian distributed affine estimation errors (Section II-A.2), which leads to the highest entropy compared with other pdfsof the same variance. Thus, it will overestimate the minimum required bit rate for finite real-world signals.

(a) *BasketballDrive* (b) *Cactus*

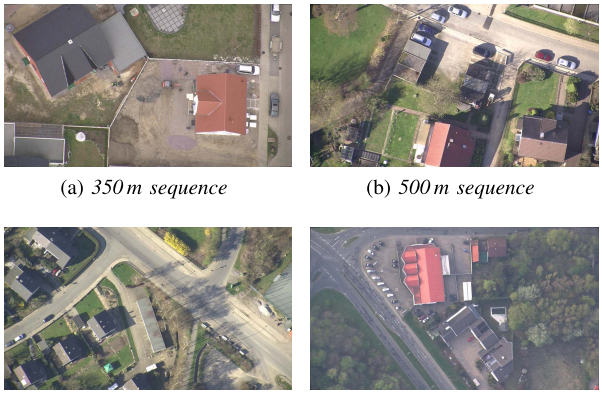
Fig. 11. JCT-VC test sequences *BasketballDrive* and *Cactus* (both full HD resolution, 50 fps) [10].

In conclusion, it has been proven that the model provides valuable indications of the prediction error bit rate as a function of the affine motion estimation accuracy. It was verified by measurements that the model qualitatively perfectly predicts the behavior of the prediction error bit rate. Due to several assumptions made in the model which approximate real-world signals, the result obtained by the model can be considered as a supremum for the minimum required bit rate for encoding the prediction error.

B. Operational rate-distortion diagrams using JEM without and with affine motion-compensated prediction

To evaluate the performance of (simplified) affine motion compensated prediction in video coding, video sequences with different characteristics are encoded using JEM 7.1 (SVN revision 603) [16] with and without affine motion compensation. From the JCT-VC test set [10], the full HD resolution sequences *BasketballDrive* and *Cactus*, both recorded at 50 fps, were arbitrarily selected to represent natural video content (Fig. 11).

Predominantly planar, high quality, full HD resolution aerial sequences with a preferentially translational global motion, recorded at 30 fps, are represented by the TNT *Aerial Video Testset* (TAVT) sequences (set 1) named *350 m sequence*, *500 m sequence*, *1000 m sequence* and *1500 m sequence* (Fig. 12) [13], [40]. The names represent the approximate recording height and indicate that for higher altitudes the ground resolution decreases, since the camera settings have not been



(a) *350 m sequence* (b) *500 m sequence*

(c) *1000 m sequence* (d) *1500 m sequence*

Fig. 12. Test sequences from the TNT *Aerial Video Testset* (TAVT) [13], [40].

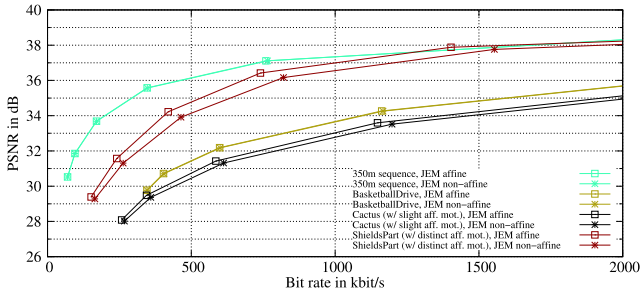


Fig. 13. Operational rate-distortion curves for different sequences encoded by JEM 7.1 (SVN revision 603) [16] with (squares) and without (stars) simplified affine motion compensation. Results for the *random access* (RA) profile are displayed. Sequences containing high amounts of non-translational motion (e. g. *ShieldsPart*) clearly profit from simplified affine motion compensation whereas sequences without such motions (e. g. *350 m sequence*, *BasketballDrive*) do not benefit from the simplified affine motion model.

changed for the sequences. As an example for a test sequence containing distinct non-translational affine motion, *ShieldsPart* is used (Fig. 10, see above) [14].

Operational rate-distortion (RD) curves for 500 frames of the test sequences each, except for *ShieldsPart* which only consists of 100 frames, are measured for the *random access* (RA) and the *low-delay p* (LDP) profile for the cases of enabled and disabled simplified affine motion-compensated prediction. The results for RA are shown in Fig. 13, whereas squares represent enabled and stars disabled simplified affine motion-compensated prediction, respectively. Identically colored curves belong to the same sequences. For reasons of clarity, the operational RD curves for the TAVT sequences are represented by only the *350 m sequence*, since the other sequences behave similarly albeit at other bit rate levels. It is obvious that for the sequence *BasketballDrive* from the JCT-VC test set as well as for the *350 m sequence* (and accordingly the other three TAVT sequences used here but not shown in the graph) only small gains can be achieved by using affine motion compensation. For the evaluation, Bjøntegaard delta (BD) rates were calculated, which measure the average bit rate difference between two rate-distortion curves [43], [44]. For the test sequences BD rate gains of only 0.88 % (LDP) and 0.54 % (RA) for *BasketballDrive*, and 0.33 % (LDP) and 0.41 % (RA) for the *350 m sequence* were achieved. However, the BD rate gains for the *Cactus* sequence including rotating elements are 6.32 % for LDP and 5.48 % for RA. For the sequence *ShieldsPart* containing a considerable amount of non-translational affine zoom motion, the observed gains even increase to 24.75 % (LDP) and 13.29 % (RA).

For sequences containing distinct non-translational affine motion, affine motion-compensated prediction may highly increase the coding efficiency of upcoming video coding standards. Especially aerial sequences captured from a drone with high amounts of rotation and zoom may highly benefit from affine motion-compensated prediction.

IV. Conclusion

In this work affine motion-compensated prediction in video coding was analyzed. The minimum bit rate required for encoding the prediction error was derived for a fully affine motion model with six degrees of freedom as well as for a simplified affine motion model with only four degrees of freedom. Both models are of particular interest since they are investigated in the standardization activities from JVET in the context of a video coding standard succeeding HEVC, likely to be named *Versatile Video Coding* (VVC). By using the rate-distortion theory, the minimum required bit rate for encoding the prediction error as a function of the motion estimation accuracy has been modeled (derivation flowchart in Fig. 1). To achieve this, the parameters of each affine motion model were assumed to be affected by statistically independent estimation errors with probability density functions (pdfs) following zero-mean Gaussian distributions. From the joint pdf of these parameter estimation errors, the pdfof the displacement estimation error in the image was derived. In contrast to previously existing models, e. g. for purely translational motion-compensated prediction, the displacement estimation error is location-dependent in the case of any affine motion-compensated prediction. The pdfof the displacement estimation error is Gaussian distributed as well and is a function of the affine parameter estimation errors. By combining the Fourier transform of the pdfof the displacement estimation error and the modeled power spectral density (PSD) of videos, the PSDof the prediction error is derived. Applying the rate-distortion theory results in the minimum required bit rate for encoding the prediction error for a given signal-to-noise ratio (SNR). Due to the Gaussian distribution assumptions and since the Gaussian distribution has the highest entropy among all distributions with same mean and variance, the supremum of the minimum bit rate required for encoding the prediction error was finally obtained.

Furthermore, the model error was determined which occurs if a translational motion model is used for a sequence containing motions, which can only be described by an affine motion model, i. e. rotation, scaling and shearing (“affinities”) in addition to translation. The results show that both, the affine parameter estimation errors as well as the affinities inherently contained in a sequence, can be mathematically modeled in the same way. For affine motion-compensated prediction to

be more efficient than simple intra coding of a HD resolution video signal itself, a bit rate of less than 2.0 bit/sample must be provided for a SNR of 30 dB. Using the example of a simplified affine motion model with a block size of 64×64 pel and a translational quarter-pel accuracy, this can only be achieved for an affine motion estimation accuracy of $\sigma_{e_a}^2 = 3 \cdot 10^{-4}$ or smaller, which can easily be achieved with real-world implementations.

Comparing the results from the fully affine motion model with those from the simplified one, it can be found that for typical affine transformation parameter estimation error variances the bit rate difference is negligible. Since the vast majority of motions contained in real-world sequences can already be described by the simplified affine motion model, only a small additional gain can be expected from a motion model additionally covering motions of very rare occurrence. Moreover, from a coding point of view, it is typically beneficial to encode as few parameters as possible. The derived model has been experimentally verified. Due to several assumptions in order to approximate the real world and since the supremum of the minimum prediction error bit rate is modeled, the absolute measured bit rates are below the modeled ones. However, in conclusion, the model provides valuable information about the minimum required motion estimation accuracy to enable a predefined bit rate for encoding the prediction error and to design upcoming video coding standards in terms of minimum required affine motion parameter accuracy.

REFERENCES

[1] *Advanced Video Coding (AVC)*, 3rd ed., document Recommendation ITU-T H.264 and ISO/IEC 14496-10 (MPEG-4 Part 10): ISO/IEC and ITU-T, Geneva, Switzerland, Jul. 2004.

[2] *ITU-T High Efficiency Video Coding (HEVC)*, 2nd ed., document Recommendation H.265/ ISO/IEC JTC 1/SC 29 23008-2:2015-05-01 MPEG-H Part 2/: ISO/IEC and ITU-T, Apr. 2015.

[3] U. Reimers, *DVB: The Family of International Standards for Digital Video Broadcasting*, 2nd ed. New York, NY, USA: Springer, 2004.

[4] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE J. Sel. Areas Commun.*, vol. SAC-5, no. 7, pp. 1140–1154, Aug. 1987.

[5] *Video Codec for Audiovisual Services at p×64 Kbit/s*, document Recommendation ITU-T H.261: ITU-T, Geneva, Switzerland, Nov. 1988.

[6] *Information Technology—Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s—Part 2: Video*, document ISO/IEC 11172-2 (MPEG-1 Part 2): ISO/IEC, Aug. 1993.

[7] *Recommendation ITU-T H.262 and ISO/IEC 13818-2 (MPEG-2 Part 2): Information Technology—Generic Coding of Moving Pictures and Associated Audio Information: ISO/IEC and ITU-T, Video*, Mar. 1995.

[8] *Video Coding for Low Bit Rate Communication*, document Recommendation ITU-T H.263: ITU-T, Geneva, Switzerland, version 1, version 2, version 3, Nov. 2000.

[9] *Information Technology—Coding of Audio-Visual Objects—Part 2: Visual*, document ISO/IEC 14496:2000-2, ISO/IEC, Dec. 2000.

[10] F. Bossen, *L1100: Common HM Test Conditions and Software Reference Configurations, Joint Collaborative Team Video Coding (JCT-VC)*, document ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29/WG 11, 12th Meeting, Geneva, Switzerland, 2013.

[11] X. Zheng, Z. Cao, and F. Wolf, *Aerial Photography Sequences for Video Coding Standard Development, Joint Video Exploration Team (JVET) of ITU-T VCEG and ISO/IEC MPEG*, document JVET-D0060, 4th Meeting, Chengdu, China, Oct. 2016. [Online]. Available: http://phenix.it-sudparis.eu/jvet/

[12] X. Zheng *et al. New Aerial Photography Sequences for Video Coding Standard Development*, document JVET-F0062, Joint Video Exploration Team (JVET) ITU-T VCEG ISO/IEC MPEG, 6th Meeting, Hobart, Australia, TAS, Mar. 2017. [Online]. Available: http://phenix.it-sudparis.eu/jvet/

[13] TNT Aerial Video Testset (TAVT). Institut für Informationsverarbeitung (TNT), Leibniz Universität Hannover. 2014. [Online]. Available: https://www.tnt.uni-hannover.de/project/TNT_Aerial_Video_Testset/

[14] L. Li *et al.*, "An efficient four-parameter affine motion model for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1934–1948, Aug. 2018.

[15] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[16] Joint Video Exploration Team (JVET). (Oct. 2017). *Joint Exploration Model (JEM)*. Accessed: Oct. 2018. [Online]. Available: https://jvet.hhi.fraunhofer.de/svn/svn_HMJEMSoftware/branches/HM-16.6-%JEM-7.1-dev/

[17] J. Chen, E. Alshina, G.-J. Sullivan, J.-R. Ohm, and J. Boyce, *Algorithm Description of Joint Exploration Test Model (JEM) 1*, Joint Video Exploration Team (JVET) ITU-T VCEG ISO/IEC MPEG, 1st Meeting: Geneva, Switzerland, JVET-A1001, Oct. 2015. [Online]. Available: http://phenix.it-sudparis.eu/jvet/

[18] H. Huang, J. W. Woods, Y. Zhao, and H. Bai, "Control-point representation and differential coding affine-motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 10, pp. 1651–1660, Oct. 2013.

[19] K. Zhang, Y.-W. Chen, L. Zhang, W.-J. Chien, and M. Karczewicz, "An improved framework of affine motion compensation in video coding," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1456–1469, Mar. 2019.

[20] C. Heithausen and J.-R. Ohm, *Inter Prediction using Estimation and Explicit Coding of Affine Parameters*, document JVET-H0031, Joint Video Exploration Team (JVET) ITU-T VCEG ISO/IEC MPEG, Macau, China, Oct. 2017.

[21] M. Koo *et al.*, *Description of SDR Video Coding Technology Proposal by LG Electronics*, document JVET-J0017, Joint Video Exploration Team (JVET) ITU-T VCEG ISO/IEC MPEG, San Diego, CA, USA, Apr. 2018.

[22] S. Parker, Y. Chen, D. Barker, P. de Rivaz, and D. Mukherjee, "Global and locally adaptive warped motion compensation in video compression," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 275–279.

[23] AOMedia Video 1 (AV1). (Jul. 2017). *AOM—AV1: How Does it Work?* [Online]. Available: https://parisvideotech.com/wp-content/uploads/2017/07/AOM-AV1-Video-Tec%h-meet-up.pdf

[24] K. Suehring and X. Li, *JVET Common Test Conditions and Software Reference Configurations, Joint Video Exploration Team (JVET) ITU-T VCEG ISO/IEC MPEG*, document JVET-H1010, 8th Meeting, Macau, China, Oct. 2017. [Online]. Available: http://phenix.it-sudparis.eu/jvet/

[25] H. Zhang, H. Chen, X. Ma, and H. Yang, *Performance Analysis of Affine Inter Prediction in JEM1.0, Joint Video Exploration Team (JVET) ITU-T VCEG ISO/IEC MPEG*, document JVET-B0037, 2nd Meeting, San Diego, CA, USA, Feb. 2016. [Online]. Available: http://phenix.it-sudparis.eu/jvet/

[26] E. Alshina, A. Alshin, K. Choi, and M. Park, *Performance of JEM 1 Tools Analysis, Joint Video Exploration Team (JVET) ITU-T VCEG ISO/IEC MPEG*, document JVET-B0022, 2nd Meeting, San Diego, CA, USA, Feb. 2016. [Online]. Available: http://phenix.it-sudparis.eu/jvet/

[27] K. Zhang, L. Zhang, H. Liu, J. Xu, Z. Deng, and Y. Wang, "Interweaved prediction for video coding," *IEEE Trans. Image Process.*, early access, Apr. 2020, doi: 10.1109/TIP.2020.2987432.

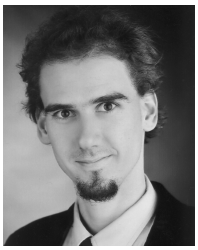
[28] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression* (Prentice-Hall Electrical Engineering Series). Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.

[29] H. Meuel, S. Ferenz, Y. Liu, and J. Ostermann, "Rate-distortion theory for affine global motion compensation in video coding," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 3593–3597.

[30] H. Meuel, *Analysis Affine Motion-Compensated Predict. its Appl. Aerial Video Coding* (Fortschritt-Berichte VDI: Reihe 10, Informatik/Kommunikation), vol. 10, no. 865. Düsseldorf, Germany: VDI Verlag GmbH, Dec. 2019.

[31] S. Klomp, *Decoderseitige Bewegungsschätzung in der Videocodierung* (Fortschritt-Berichte VDI: Reihe 10, Informatik/Kommunikation), vol. 10, no. 820. Düsseldorf, Germany: VDI Verlag GmbH, 2012.

[32] B. Bross, *Versatile Video Coding (Draft 8), Joint Video Experts Team (JVET)*, document JVET-Q2001, ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, 17th Meeting Brussels, Belgium, Feb. 2020. [Online]. Available: http://phenix.it-sudparis.eu/jvet/
[33] J. O'Neal and T. Natarajan, "Coding Isotropic Images," *IEEE Trans. Inf. Theory*, vol. IT-23, no. 6, pp. 697–707, Nov. 1977.
[34] P. Z. Peebles, *Probability, Random Variables and Random Signal Principles*. New York, NY, USA: McGraw-Hill, 1993.
[35] H.-G. Musmann, *Statistische Methoden der Nachrichtentechnik* (Lecture Notes). Hannover, Germany: Institut für Informationsverarbeitung, Leibniz Universität Hannover, May 2017.
[36] K. Pearson, "Notes regression inheritance case two parents," in *Proceedings of the Royal Society of London (Great Britain)*, vol. 58. New York, NY, USA: Taylor & Francis, Jun. 1895.
[37] S. Birchfield. (2007). *KLT: An Implementation of the Kanade-Lucas-Tomasi Feature Tracker*. [Online]. Available: https://cecas.clemson.edu/~stb/klt/
[38] Landesamt für Geoinformation und Landesvermessung Niedersachsen (LGN). (Jul. 2013). *Test Image: Hannover.Hannover. Name: dop20_32550000_5802000_col.tif, Ground Resolution: 0.2 m×0.2 m, Format: TIFF Image, File Size: 301 MB, Format: Raw, Width: 10000 pel, Height: 10000 Pel, Color Space: RGB, Bit Depth: 8 Bit*. [Online]. Available: https://www.lgln.niedersachsen.de/
[39] C. E. Duchon, "Lanczos filtering in one and two dimensions," *J. Appl. Meteorol.*, vol. 18, no. 8, pp. 1016–1022, Aug. 1979.
[40] H. Meuel, M. Munderloh, M. Reso, and J. Ostermann, "Mesh-based piecewise planar motion compensation and optical flow clustering for ROI coding," in *Proc. APSIPA Trans. Signal Inf. Process.*, vol. 4, 2015, pp. 1–19. [Online]. Available: http://journals.cambridge.org/article_S2048770315000128, doi: 10.1017/ATSIP.2015.12.
[41] X. Li, J. Boyce, P. Onno, and Y. Ye, *Common Test Conditions and Software Reference Configurations for the Scalable Test Model, Joint Collaborative Team Video Coding (JCT-VC)*, document L1009 ITU-T SG 16 WP 3 ISO/IEC JTC 1/SC 29/WG 11. 12th Meeting, Geneva, Switzerland, 2013.
[42] T. Laude, Y. G. Adhisantoso, J. Voges, M. Munderloh, and J. Ostermann, "A comprehensive video codec comparison," in *Proc. APSIPA Trans. Signal Inf. Process.*, vol. 8, Oct. 2019, pp. 1–16, doi: 10.1017/ATSIP.2019.23.
[43] G. Bjöntegaard, *Calculation of Average PSNR Differences Between RD Curves*, document VCEG-M33 ITU-T SG 16/Q6, 13th Meeting, Austin, TX, USA, Apr. 2001.
[44] G. Bjöntegaard, *Improvements of the BD-PSNR model*, document VCEG-AI11 ITU-T SG 16/Q6, 35th Meeting, Berlin, Germany, 2008.



Holger Meuel received the degree in electrical engineering and the Dipl.-Ing. degree from Technische Universität (TU) Braunschweig, Germany, in 2010, and the Dr.-Ing. (Ph.D.) degree from Leibniz Universität Hannover (LUH), Germany, in 2019. He joined Institut für Informationsverarbeitung (TNT), LUH, as a Research and Teaching Assistant, where he was the Senior Engineer from 2010 to 2017. He attended several standardization meetings for the video coding standard *High Efficiency Video Coding* (HEVC) of the MPEG and VCEG *Joint Collaborative Team on Video Coding* (JCT-VC). In that context, he also dealt with radial camera lens distortion compensation, scalable video coding, and screen content coding. He is author of more than 20 scientific publications. He filed three patent applications. His research interests are video coding with special focus on low-bit rate video coding for aerial surveillance applications.

Jörn Ostermann received the degree in electrical engineering and communications engineering from the University of Hannover and Imperial College London, and the Dipl.-Ing. and Dr.-Ing. degrees from the University of Hannover in 1988 and 1994, respectively. In 1994, he joined AT&T Bell Labs. From 1996 to 2003, he was with AT&T Labs–Research. Since 2003, he has been a Full Professor and the Head of the Institut für Informationsverarbeitung, Leibniz Universität Hannover, Germany. Since 2008, he has been the Chair of the Requirements Group, MPEG (ISO/IEC JTC1 SC29 WG11). He has published more than 100 research articles and book chapters. He is a coauthor of a graduate level text book on *video communications*. He holds more than 30 patents. His current research interests are video coding and streaming, computer vision, machine learning, 3D modeling, face animation, and computer–human interfaces. He received several international awards.