RATE-DISTORTION THEORY FOR AFFINE GLOBAL MOTION COMPENSATION IN VIDEO CODING

Holger Meuel, Stephan Ferenz, Yiqun Liu, Jörn Ostermann

Institut für Informationsverarbeitung, Leibniz Universität Hannover, Germany
Email: *{meuel, ferenz, liuyiqun, office}*@tnt.uni-hannover.de

ABSTRACT

In this work, we derive the rate-distortion function for video coding using affine global motion compensation. We model the displacement estimation error during motion estimation and obtain the bit rate after applying the rate-distortion theory. We assume that the displacement estimation error is caused by a perturbed affine transformation. The 6 affine transformation parameters are assumed statistically independent, with each of them having a zero-mean Gaussian distributed estimation error. Based on that, the joint p.d.f. of the displacement estimation errors is derived and related to the prediction error. Using the rate-distortion theory, we calculate the bit rate in dependence of the perturbation of the affine transformation parameters. Comparing with a translational motion model in video coding standards like HEVC, we determine accuracy boundaries for the affine transformation, with which a gain can be achieved.

Index Terms— Affine Transformations, Global Motion Compensation, GMC, Rate-Distortion Theory, Efficiency Analysis, Aerial Video Coding, ROI Coding

1. INTRODUCTION

Motion compensated prediction (MCP) is one of the key elements in modern hybrid video coding standards like *Advanced Video coding* (AVC) [1] or *High Efficiency Video Coding* (HEVC) [2]. MCP was and is typically performed block-wise for blocks of different sizes, *e. g.* of 4×4 up to $16\times16\,\text{pel}^2$ for AVC or $64\times64\,\text{pel}^2$ for HEVC. The minimum bit rate of the prediction error of motion compensated prediction in dependence of the variance of the displacement estimation error was theoretically analyzed by Girod [3].

For video sequences with distinct global motion, (affine) global motion compensation (GMC) was introduced in *MPEG-4 Advanced Simple Profile (MPEG-4 ASP)* [4]. Since the coding efficiency gains of GMC stayed behind the expectation for general video coding for natural scenes without prevalent global motion, GMC was removed in the MPEG-4 ASP successor AVC again and was replaced by an improved *Motion Vector Prediction (MVP)*.

Nowadays, new scenarios with distinct global motion—like videos captured from *Unmanned Aerial Vehicles (UAV)*/*Micro Aerial Vehicles (MAV)* like multicopters—emerge and are also considered in recent test sets [5, 6, 7]. These video sequences contain higher order global motion, which cannot accurately be described by a purely translational motion model, *e. g.* caused by a tilted camera. To cope with such motions better, the *ITU-T/ISO/IEC Joint Video Exploration Team (JEVT)* (on Future Video Coding) incorporated a (simplified 4-parameter) affine motion model into their reference software *Joint Exploration Model (JEM)* [8] again [9], whereas in contrast to MPEG-4 ASP, it works on block-level. Affine (as well as homographic) global motion compensation is also contained in the video codec AV1 [10]. Early JVET studies based on the initial JEM software (ver. 1.0) on the common test set [11] (containing no such sequences) show coding efficiency gains of up to 1.35 % (JEM 1.0, configuration *Low Delay P (LDP) main 10*) [12] which is the 7th best of 22 of all proposals for next generation video encoding till then [13]. Larger gains can be expected for sequences containing more higher order motions like rotation or zoom [14, 15].

Although affine global motion compensation has a long tradition in video coding, it has not been theoretically analyzed in the context of video coding.

In this work, we present an efficiency analysis of affine motion compensation. We analytically derive the prediction error after motion compensation in dependence of the affine transform parameter accuracy in Section 2. Using the rate-distortion theory [16], we derive the bit rate from the prediction error in Subsection 2.1. Results for a typical signal-to-noise ratio (SNR) will be presented in the simulations in Section 3 using the example of aerial video sequences containing distinct global motion. Section 4 finally concludes the paper.

2. EFFICIENCY ANALYSIS OF AFFINE MOTION COMPENSATION IN VIDEO CODING

Assuming a full affine motion model with 6 parameters, we can compute the coordinates x and y in the current (destination) frame from the affine parameter matrix $\mathbf{A} = \left(\begin{smallmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{smallmatrix}\right)$ and the homogeneous coordinates $(x', y', 1)$ in the source frame

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

The parameters a_{13} and a_{23} describe the translational part of a motion, whereas the parameters $a_{11,12,21,22}$ express the rotation, scaling and shearing. These 4 parameters are further referred as (purely) “affine parameters”. We assume that each parameter a_{ij} with $i = \{1,2\},\ j = \{1,2,3\}$ is perturbed by an independent error term e_{ij}, caused by inaccurate parameter estimation. Consequently, the perturbed $\hat{x}$ coordinate can be expressed as $\hat{x} = \hat{a}_{11}x' + \hat{a}_{12}y' + \hat{a}_{13}$, leading to estimation errors in horizontal and vertical direction of Δx and Δy (in pel)

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

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

Assuming each error term e_{ij} to be zero-mean Gaussian distributed leads to the probability density functions (p.d.f.s)

$$p(e_{ij}) = \frac{1}{\sqrt{2\pi\sigma^2_{e_{ij}}}}\cdot\exp\left(-\frac{e^2_{ij}}{2\sigma^2_{e_{ij}}}\right) \tag{4}$$

with $i = \{1, 2\}$ and $j = \{1, 2, 3\}$.

We assume a Gaussian distribution as the worst-case scenario since it has the maximal entropy of all distributions with the same variance. Moreover, the affine parameter estimation is typically based on a high number of feature point correspondences, with each having an independently distributed subpel error. Thus, our Gaussian assumption is additionally justified by the central limit theorem. For statistically independent variables we get a joint p.d.f. $p_{E_{11},\dots,E_{23}}(e_{11}, \dots, e_{23})$ for the random variables $E_{11}, \dots, E_{23}$ generating the observations $e_{11}, \dots, e_{23}$:

$$p_{E_{11},\dots,E_{23}}(e_{11}, \dots, e_{23}) = p(e_{11}) \cdot \ \dots \ \cdot p(e_{23}) \ . \tag{5}$$

To convert the p.d.f. $p_{E_{11},\dots,E_{23}}(e_{11}, \dots, e_{23})$ to the desired p.d.f. $p_{\Delta X,\Delta Y}(\Delta x, \Delta y)$ of the resulting pixel errors Δx, Δy caused by affine parameter inaccuracies, we use the transformation theorem for p.d.f.s ([17, 18])

$$p_{\mathscr{Y}_1,\dots,\mathscr{Y}_M}(\mathscr{y}_1, \dots, \mathscr{y}_M) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p_{\mathscr{X}_1,\dots,\mathscr{X}_N}(\xi_1, \dots, \xi_N)$$
$$\cdot \prod_{m=1}^{M} \delta(\mathscr{y}_m - g_m(\xi_1, \dots, \xi_N)) \mathrm{d}\xi_1 \dots \mathrm{d}\xi_N \ , \tag{6}$$

with $\delta(\cdot)$ denoting the Dirac delta function, $g_1, \dots, g_M$ being functions $\mathscr{y}_1 = g_1(\mathscr{x}_1, \dots, \mathscr{x}_N)$, $\dots$, $\mathscr{y}_M = g_M(\mathscr{x}_1, \dots, \mathscr{x}_N)$ and $p_{\mathscr{Y}_1,\dots,\mathscr{Y}_M}(\mathscr{y}_1, \dots, \mathscr{y}_M)$ being the compound p.d.f. With equations (2) and (3) this yields

$$\begin{aligned} p_{\Delta X,\Delta Y}(\Delta x, \Delta y) = &\int_{\mathbb{R}^6} p_{E_{11},\dots,E_{23}}(e_{11}, \dots, e_{23}) \\ &\cdot \delta(\,\Delta x - (x' e_{11} + y' e_{12} + e_{13})\,) \\ &\cdot \delta(\,\Delta y - (x' e_{21} + y' e_{22} + e_{23})\,) \, \mathrm{d}e_{11} \dots \mathrm{d}e_{23} \ , \end{aligned} \tag{7}$$

with a dependency on the location coordinates x', y' in the source frame.

By using the properties of the delta function, we solve two integrals

$$\begin{aligned} &p_{\Delta X,\Delta Y}(\Delta x, \Delta y) \\ &= \int_{\mathbb{R}^4} p_{E_{11},\dots,E_{22}}(e_{11}, e_{12}, \Delta x - x' e_{11} - y' e_{12}, e_{21}, e_{22}, \\ &\quad \Delta y - x' e_{21} - y' e_{22}) \, \mathrm{d}e_{11} \mathrm{d}e_{12} \mathrm{d}e_{21} \mathrm{d}e_{22} \ . \end{aligned} \tag{8}$$

Exploiting the statistical independence (equation (5)), we separate the integrands, which leads to

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

The following derivation is presented only for the x component, since the y component can be calculated similarly. For Δx we get from equation (9) with equation (4)

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

Integration results in

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

After calculating the y component accordingly, we obtain the resulting displacement estimation error

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

$$\text{with } \sigma^2_{\Delta x} = \sigma^2_{e_{11}} x'^2 + \sigma^2_{e_{12}} y'^2 + \sigma^2_{e_{13}} \tag{13}$$

$$\text{and } \sigma^2_{\Delta y} = \sigma^2_{e_{21}} x'^2 + \sigma^2_{e_{22}} y'^2 + \sigma^2_{e_{23}} \ . \tag{14}$$

As can be seen, the variances $\sigma^2_{\Delta x}$ and $\sigma^2_{\Delta y}$ depend on the locations x', y'.

2.1. Rate-distortion analysis of affine global motion compensated prediction

To derive the bit rate for coding the prediction error in motion compensated video coding, we use the findings from Girod, who related the displacement estimation error $p_{\Delta X,\Delta Y}(\Delta x, \Delta y)$ to the prediction error e_p [3]. Applying the rate-distortion theory [16] results in the minimum achievable bit rate for encoding the prediction error. In this subsection we will summarize the derivations from [3].

Given a displacement estimation error $p_{\Delta X,\Delta Y}(\Delta x, \Delta y)$, we obtain the power spectral density of the prediction error

$$S_{ee}(\Lambda) = 2\, S_{ss}(\Lambda)\, [1 - \mathrm{Re}(P(\Lambda)] + \Theta \, , \tag{15}$$

where $S_{ss}(\Lambda)$ denotes the power spectral density of the video signal s, Λ being the two-dimensional (2D) spatial frequency vector $\Lambda := (\omega_x, \omega_y)$, $P(\Lambda)$ being the 2D Fourier transform of the probability density function (p.d.f.) of the displacement estimation error, and Θ being a parameter that generates the function $R(D)$ by taking on all positive real values ([3], equation (28)). The power spectral density $S_{ss}(\omega_x, \omega_y)$ was determined according to O'Neil and Girod [19, 3]. There it was assumed that the statistics of the source can be

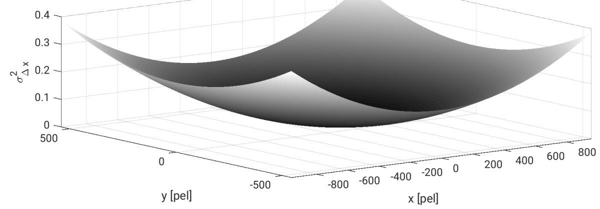


Fig. 1. Location dependent variance $\sigma^2_{\Delta x}$ of Gaussian distributed displacement estimation error p.d.f.s for an HD image, $\sigma^2_{e_{11}} = 2.3\text{e}-7$, $\sigma^2_{e_{12}} = 4.6\text{e}-7$ [7, 21] and $\sigma^2_{e_{13}} = 0.04$ [3].

represented by the autocorrelation function

$$R_{ss}(\Delta x, \Delta y) = E\left[s(x, y) \cdot s(x - \Delta x, y - \Delta y)\right]$$
$$:= \exp\left(-\alpha \sqrt{\Delta x^2 + \Delta y^2}\right) . \tag{16}$$

We assume α not to be isotropic and thus replace it by $\alpha := \sqrt{\alpha_x \alpha_y}$. 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 pixels $\alpha_x = -\ln(\rho_x)$ and $\alpha_y = -\ln(\rho_y)$ [19]. For this, the Pearson correlation coefficients $\rho(X, Y) = \frac{\text{cov(X,Y)}}{\sigma_X \sigma_Y}$ and similarly ρ_Y with the standard deviations σ_X, σ_Y and the covariance cov were determined [20]. The desired power spectral density $S_{ss}(\Lambda)$ to be inserted in equation (15) is now the Fourier transform of equation (16).

Finally, we derive the distortion D as well as the corresponding minimum transfer rate $R(D)$ from the rate-distortion function for a given mean-squared error ([3], equations (19–20))

$$D = \frac{1}{4\pi^2} \iint\limits_{\Lambda} \min\left[\Theta, S_{ss}(\Lambda)\right] \mathrm{d}\Lambda , \tag{17}$$

$$R(D) = \frac{1}{8\pi^2} \iint\limits_{\substack{\Lambda : (S_{ss}(\Lambda) > \Theta \\ \text{and } S_{ee}(\Lambda) > \Theta)}} \log_2\left[\frac{S_{ee}(\Lambda)}{\Theta}\right] \mathrm{d}\Lambda \text{ bit} . \tag{18}$$

We would like to emphasize that our $\sigma^2_{\Delta x}$ and $\sigma^2_{\Delta y}$ are location dependent, since they are functions of the source pixel coordinates x', y'. Consequently, $p_{\Delta X, \Delta Y}(\Delta x, \Delta y)$, $P(\Lambda)$ and $S_{ee}(\Lambda)$ are also location dependent.

Using the idea of generating the rate-distortion function for translative motion like explained by Girod [3] and our results from Section 2, we derived the rate-distortion function for affine motion.

3. SIMULATIONS

In our simulations, we evaluate the minimal bit rate for affine global motion compensated prediction.

As we have seen in the last section, the variances of the displacement estimation error $p_{\Delta X, \Delta Y}(\Delta x, \Delta y)$ depends on the location in the image according to equations (13) and (14). Thus, also the resulting minimum achievable bit rate is location dependent. To obtain the total bit rate for encoding one frame, we summarize the pel-wise bit rates afterwards.

Table 1. Measured variances $\sigma^2_{e_{ij}}$ of affine transformation parameters of aerial videos from the TAVT data set [7, 21].

	$\sigma^2_{e_{11}}$	$\sigma^2_{e_{12}}$	$\sigma^2_{e_{21}}$	$\sigma^2_{e_{22}}$	mean $(\sigma^2_{e_{11}}, \sigma^2_{e_{22}})$	mean $(\sigma^2_{e_{12}}, \sigma^2_{e_{21}})$
350m seq.	2.03e−7	6.03e−7	6.59e−7	2.24e−7	**2.13e−7**	**6.31e−7**
500m seq.	1.94e−7	5.09e−7	3.63e−7	1.94e−7	**1.94e−7**	**4.35e−7**
1000m seq.	1.74e−7	4.05e−7	4.13e−7	2.12e−7	**1.93e−7**	**4.09e−7**
1500m 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**

Table 2. Measured horizontal and vertical correlations between adjacent pixels for typical test sequences (*: 100 frames each).

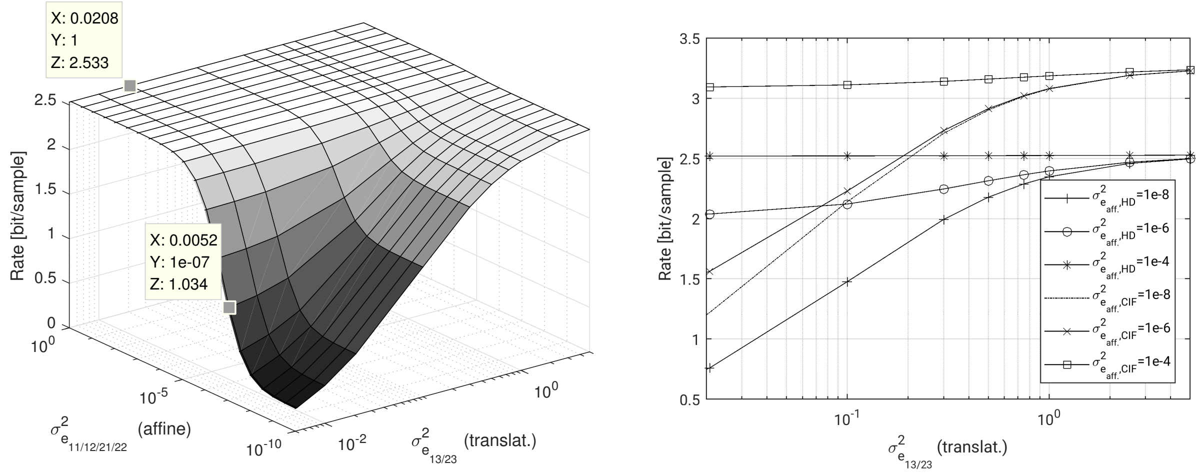
Sequence	Corr. ρ_x	Corr. ρ_y
Values from Girod [3]	0.928	0.934
BasketballDrive* (HD) [25]	0.9782	0.9488
BQTerrace* (HD) [25]	0.9680	0.9659
Cactus* (HD) [25]	0.9741	0.9812
Kimono* (HD) [25]	0.9883	0.9900
ParkScene* (HD) [25]	0.9634	0.9518
Mean of CIF seq. Claire, Foreman, Mobile*[24]	0.9402	0.8958
Mean of HD sequences* [25]	**0.9744**	**0.9677**

Measured variances using an affine motion estimation based on a KLT feature tracker [22] and RANSAC [23] for the aerial video sequences from the TAVT data set [7, 21] are given in Table 1, assuming that no non-translational motion is prevalent between two consecutive frames. From these results it is obvious that the variances $\sigma_{e_{11}}$ and $\sigma_{e_{22}}$ as well as $\sigma_{e_{12}}$ and $\sigma_{e_{21}}$ are very similar. This can be explained, if we consider the rotational part of the affine transform to be caused by a physical rotation of the camera and the skew-symmetry of a 2D rotation matrix. Justified by our findings, we assume $\sigma_{e_{11}} = \sigma_{e_{22}}$ and $\sigma_{e_{12}} = \sigma_{e_{21}}$ and use the average values (see Table 1). For illustration, the location dependent variance $\sigma^2_{\Delta x}$ is shown in Fig. 1 for a full HD resolution image of $1920 \times 1080\,\text{pel}^2$ and variances like observed in the TAVT data set [7].

For calculating the power spectral density S_{ss} of the video signal in equation (15) and the distortion in equation (17), we determine the exponential drop rates α_x and α_y of the autocorrelation function (equation (16)). We measured the mean correlation of horizontally and vertically adjacent pixels of several video sequences. To demonstrate the effect of different resolutions, we use CIF sequences (352×288) [24] as well as full HD resolution sequences (1920×1080) from the JCT-VC test set [25]. Results can be found in Table 2.

Evaluation of the rate-distortion theory results in minimum required bit rates for different variances $\sigma^2_{e_{ij}}$ of Gaussian displacement estimation error p.d.f.s for a distortion of SNR = 30 dB of the affine transform parameters in Fig. 2. For the simulations we assumed all affine parameters to be equal ($\sigma^2_{e_{11}} = \sigma^2_{e_{12}} = \sigma^2_{e_{21}} = \sigma^2_{e_{22}}$) as well as both translational parameters ($\sigma^2_{e_{13}} = \sigma^2_{e_{23}}$).

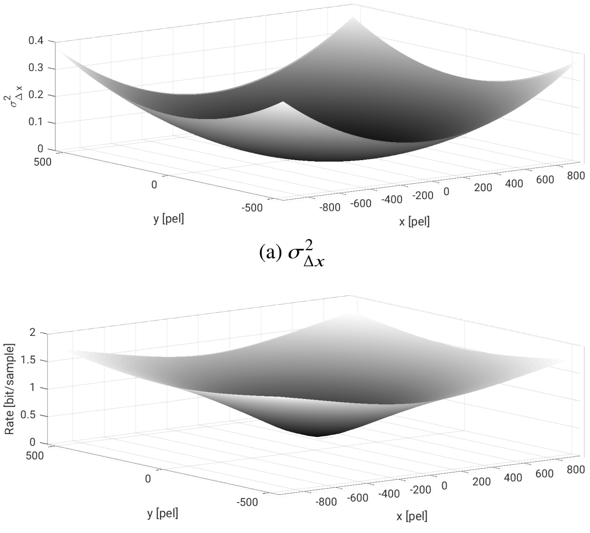
In a second experiment, we calculated the displacement vector field for several simulated affine transform matrices ($N = 100$). The affine matrices were assumed to reflect rotation, scaling and shearing motion by having Gaussian distributed parameters a_{11}, a_{22} with a mean value of 1 and variances of $\sigma^2_{a_{11}} = \sigma^2_{a_{22}} = 2.3\text{e–}7$ and parameters a_{12}, a_{21} with a mean value of 0 and variances of $\sigma^2_{a_{12}} = \sigma^2_{a_{21}} = 4.6\text{e–}7$ (see Table 1). This corresponds to a rotational error of about 0.20 and 0.23 degree, respectively. The location dependent variance of the displacement vector field is shown in Fig. 3a. The results fit to our derivations in Section 3, assuming the variances of the parameters $a_{11}, \ldots, a_{22}$ being the variances of the errors $e_{11}, \ldots, e_{22}$. The location dependent variance $\sigma^2_{\Delta_x}$ are marginally smaller than



(a) For HD resolution. Isolines for corresponding quarter-pel resolution ($\sigma^2_{e_{13/23}} = 0.0052$) and half-pel resolution of the translational motion estimation accuracy are marked by datatips.

(b) 2D cuts of surface (Fig. 2a) for HD and CIF sequences.

Fig. 2. Minimum required bit rate versus variances $\sigma^2_{e_{ij}}$ of Gaussian displacement estimation error p.d.f.s for a distortion of SNR = 30 dB assuming $\sigma^2_{e_{11}} = \sigma^2_{e_{12}} = \sigma^2_{e_{21}} = \sigma^2_{e_{22}}$ and $\sigma^2_{e_{13}} = \sigma^2_{e_{23}}$. The surface left (a) is for HD resolution, the 2D cuts right (b) contain plots for HD and CIF resolution and accordingly different correlations.



(a) $\sigma^2_{\Delta x}$

(b) Bit rate

Fig. 3. Simulated location dependent variance $\sigma^2_{\Delta x}$ (a) and bit rate (b) for Gaussian distributed displacement estimation error p.d.f.s for a HD image for given variances $\sigma^2_{e_{11}} = \sigma^2_{e_{22}} = 2.3\text{e–}7$, $\sigma^2_{e_{12}} = \sigma^2_{e_{21}} = 4.6\text{e–}7$.

the calculated ones in Fig. 1, since in the simulation in Fig. 3a we assumed the translational error to be zero ($\sigma^2_{e_{13}} = \sigma^2_{e_{23}} = 0$).

From the displacement vector field, the location dependent bit rates are derived according to Section 2. They are shown Fig. 3b.

From our results, we infer:

1. The variance of the displacement estimation error of the purely affine parameters ($\sigma^2_{e_{11/12/21/22}}$) has to be magnitudes smaller than the variance of the translational parameters ($\sigma^2_{e_{13/23}}$). This can be considered as realistic (see Table 1), since the estimation accuracy of the pure affine parameters is not limited to a specific fractional-pel motion vector accuracy.
2. Assuming a sequence with a specific degree of purely affine motion (we call it "affinity"), which cannot be described by a translational motion model, the minimum bit rate is limited along the affine-variances-axis (directing from the origin to the left in Fig. 2a). As an example, we assume a HD sequence with an "affinity" of 1e−7. Then, the minimum bit rate for encoding the prediction error using a translational motion estimator with the very small displacement estimation error variance of $\sigma^2_{e_{13}} = \sigma^2_{e_{23}} = 0.0052$ (which equals 1/4 pel resolution) is 1.034 bit/sample. In contrast to that the minimum bit rate is only 0.039 bit/sample for an accurate *affine* motion estimator with $\sigma^2_{\text{affine}} = 1\text{e}-8$.
3. From the example we generalize that the minimum required bit rate is reached, if the motion model covers the real motion contained in the scene, *and* if the variance of the estimator is smaller than the "affinity" contained in the scene.
4. As it is obvious from equations (12)–(14) (and Fig. 1), $\sigma^2_{\Delta x}$ and $\sigma^2_{\Delta y}$ grow for large image dimensions. For block-based motion compensation, the "image dimensions" are equal to the block dimensions. Thus, the gain introduced by affine *block-based* motion compensation may be much more insignificant.

4. CONCLUSION

In our paper we derive the minimum required bit rate for encoding the prediction error of affine (global) motion compensated prediction by applying the rate-distortion theory. We derived accuracy requirements for the affine parameter estimation for which an affine motion model is beneficial in terms of coding efficiency. Scenes, which contain high degrees of purely affine motion (i. e. rotation, scaling, shearing), can be described much better. Consequently, the working point moves towards much smaller bit rates resulting in higher encoding efficiency. Considering the location dependency of the displacement estimation error for affine global motion, the employment of affine motion compensation on block-level remains questionable.

5. REFERENCES

[1] AVC, *Rec. ITU-T H.264 & ISO/IEC 14496-10 MPEG-4 Pt.10: Adv. Video Cod. (AVC)-3rd Ed.*, ISO/IEC&ITU-T, Geneva, Switzerland, July 2004.

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

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

[4] ISO/IEC, *ISO/IEC 14496:2000-2: Information technology - Coding of Audio-Visual Objects - Part 2: Visual*, Dec. 2000.

[5] 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, 4th Meeting: Chengdu, CN, from Oct. 15–21, 2016, Document: JVET-D0060*, Oct 2016.

[6] X. Zheng, W. Li, Z. Cao, W. Su, C. Zhao, Y. Li, Z. Lorenz, H. Wu, Z. Du, and D. A. Hoang, "New aerial photography sequences for video coding standard development," *Joint Video Exploration Team (JVET) of ITU-T VCEG and ISO/IEC MPEG, 6th Meeting: Hobart, AU, from Mar. 31–April 07, 2017, Document: JVET-F0062*, Mar 2017.

[7] Institut für Informationsverarbeitung (TNT), Leibniz Universität Hannover, "TNT Aerial Video Testset (TAVT)," 2010–2014, https://www.tnt.uni-hannover.de/project/TNT_Aerial_Video_Testset/.

[8] JVET, "Joint Exploration Model (JEM) of the Joint Video Exploration Team (on Future Video coding) of ITU-T VCEG and ISO/IEC MPEG (JVET)," .

[9] 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) of ITU-T VCEG and ISO/IEC MPEG, 1st Meeting: Geneva, CH, from Oct. 19–21, 2015, Document: JVET-A1001*, Oct 2015.

[10] AOMedia Video 1 (AV1), "AOM – AV1: How does it work?," July 2017, https://parisvideotech.com/wp-content/uploads/2017/07/AOM-AV1-Video-Tech-meet-up.pdf.

[11] K. Suehring and X. Li, "JVET common test conditions and software reference configurations," *Joint Video Exploration Team (JVET) of ITU-T VCEG and ISO/IEC MPEG, 2nd Meeting: San Diego, US, from Feb. 22–26, 2016, Document: JVET-B1010*, Feb 2016.

[12] H. Zhang, H. Chen, X. Ma, and H. Yang, "Performance analysis of affine inter prediction in JEM1.0," *Joint Video Exploration Team (JVET) of ITU-T VCEG and ISO/IEC MPEG, 2nd Meeting: San Diego, US, from Feb. 20–26, 2016, Document: JVET-B0037*, Feb 2016.

[13] E. Alshina, A. Alshin, K. Choi, and M. Park, "Performance of JEM 1 tools analysis," *Joint Video Exploration Team (JVET) of ITU-T VCEG and ISO/IEC MPEG, 2nd Meeting: San Diego, US, from Feb. 20–26, 2016, Document: JVET-B0022*, Feb 2016.

[14] Li Li, Houqiang Li, Dong Liu, Zhu Li, Haitao Yang, Sixin Lin, Huanbang Chen, and Feng Wu, "An Efficient Four-Parameter Affine Motion Model for Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[15] Li Li, Houqiang Li, Zhuoyi Lv, and Haitao Yang, "An affine motion compensation framework for high efficiency video coding," in *IEEE Internat. Symp. on Circuits and Systems (ISCAS)*, May 2015, pp. 525–528.

[16] Toby Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice-Hall electrical engineering series. Prentice-Hall, 1971.

[17] Peyton Z. Peebles, Jr., *Probability, random variables and random signal principles*, McGraw-Hill Education, New York, 1993, English.

[18] Hans-Georg Musmann, "Statistische Methoden der Nachrichtentechnik," Inst. für Informationsverarbeitung, Leibniz Universität Hannover, May 2017, Lecture Notes.

[19] J. O'Neal and T. Natarajan, "Coding Isotropic Images," *IEEE Transactions on Information Theory*, vol. 23, no. 6, pp. 697–707, Nov. 1977.

[20] Royal Society (Great Britain), *Proceedings of the Royal Society of London*, Number Bd. 58. Taylor & Francis, 1895.

[21] Holger Meuel, Marco Munderloh, Matthias Reso, and Jörn Ostermann, "Mesh-based Piecewise Planar Motion Compensation and Optical Flow Clustering for ROI Coding," in *APSIPA Transact. on Sig. and Inform. Proc.*, 2015, vol. 4.

[22] Stan Birchfield, "KLT: An Implementation of the Kanade-Lucas-Tomasi Feature Tracker," 2007.

[23] Martin A. Fischler and Robert C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, June 1981.

[24] "Test sequences *Foreman*, *Claire*, *Mobile*," https://media.xiph.org/video/derf/.

[25] Xiang Li, Jill Boyce, Patrice Onno, and Yan Ye, "L1009: Common Test Conditions and Software Reference Configurations for the Scalable Test Model. Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11. 12th Meeting, Geneva, CH, 14-23 Jan," 2013.