Region-based models for motion compensation in video compression

Jean Bégaint*†, Franck Galpin*, Philippe Guillotel* and Christine Guillemot†

*Technicolor
Av. des Champs Blancs
35576 Cesson-Sévigné, France
firstname.lastname@technicolor.com

† INRIA
Campus de Beaulieu
35042 Rennes, France
firstname.lastname@inria.fr

***Abstract*—Video codecs are primarily designed assuming that rigid, block-based, two-dimensional displacements are suitable models to describe the motion taking place in a scene. However, translational models are not sufficient to handle real world motion types such as camera zoom, shake, pan, shearing or changes in aspect ratio. We present here a region-based inter-prediction scheme to compensate such motion. The proposed mode is able to estimate multiple homography models in order to predict complex scene motion. We also introduce an affine photometric correction to each geometric model. Experiments on targeted sequences with complex motion demonstrate the efficiency of the proposed approach compared to the state-of-the-art HEVC video codec.**

I. Introduction

The efficiency of video compression tools heavily relies on their ability to reduce the temporal redundancy between consecutive frames. The classical and current approach consists in predicting the motion between frames by estimating translational motion vectors. A prediction is then performed by translating blocks from the reference frame according to the corresponding motion vectors. A potential residue is added when this prediction is inaccurate in respect to the targeted encoding bit-rate or quality.

However, classical translational models cannot compensate accurately some specific motion types such as camera zoom, shake, shearing, pan, changes in aspect ratio. Such complex motion are currently handled by splitting large objects into multiple coding blocks compensated with translational motion models. This requires more side information to code the block splitting tree and produces inaccurate predictions, which consequently result in costly residues.

Using more complex transformation models has long been investigated by the video compression community. Early attempts were proposed for MPEG-4 [1] to apply homographic global motion compensation to sprites [2], [3] and for an associated global and local compensation [4] in H.263 [5]. These approaches were discarded at that time in favor of translational models and dense block partitioning, both for coding performances and complexity reasons. Recent works have demonstrated that coding improvements could still be achieved by using global homographic motion models in current state-of-the-art video codecs.

In the ongoing work to improve the compression efficiency of the High Efficiency Video Coding (HEVC) [6] codec, an affine mode was proposed in [7] to the Joint Video Exploration Team (JVET) [8]. Chen *et al.* proposed a simplified affine prediction model estimated at the block level which achieves significant gains on the targeted sequences and good results on the common test condition (CTC) sequences [9].

Recently, support for global (frame-based) and local (block-based) homography models has also been proposed and integrated in the emerging AV1 codec from the Alliance for Open Media [10], [11]. Parker *et al.* demonstrated the effectiveness of such motion prediction tools on videos with complex or steep camera motion. However, they noted that a single global motion model may not be sufficient to handle geometric distortions in scenes with strong parallax.

In a previous work [12], the authors proposed a novel prediction method for compressing highly correlated images as found in photo albums or image cloud databases. In order to compensate for strong distortions such as differences of viewpoint, focal length or cameras, a region-based compensation method was developed. We adapt here the previously proposed scheme to be used for video compression in a classical state-of-the-art video codec (HEVC).

In this paper, we present a region-based prediction mode for motion compensation in video compression. The proposed approach aims at extracting multiple homographic transformation models between video frames. A photometric compensation model is also estimated for each segmented region. We demonstrate the efficiency of the proposed scheme on sequences with non-translational motion, compared both to the classical translational compensation and to a single global homographic motion model. Coding results were obtained with the reference HEVC implementation (HM software[1]).

II. Region-based models estimation

A. Single homography model estimation

To model large geometric distortion between frames, a homography model is often used as it can handles many types

[1] https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/

of distortions. A homography model H can be defined by the following 3×3 matrix

$$H = \begin{bmatrix} s_x.\cos(\theta) & -s_y.\sin(\theta+\sigma) & t_x \\ s_x.\sin(\theta) & s_y.\cos(\theta+\sigma) & t_y \\ k_x & k_y & 1 \end{bmatrix} \quad (1)$$

where (t_x, t_y) represent the translation coefficients, θ the rotation, (s_x, s_y) the scale parameters, σ the shear, and (k_x, k_y) the keystone distortion coefficients. To estimate the homography parameters, the first step consists in matching common landmarks between the frames. The model parameters are then estimated from the matched locations by minimizing the projection error between the projected keypoints of the reference frame and the target frame keypoints.

To detect and match common landmarks, local feature descriptors are often used as they are more robust to geometric distortions (*e.g.* translation, rotation, zoom, scale) and illumination variations than pixel values [13]. These keypoint descriptors are used to detect and match common landmarks between frames. Feature vectors (or descriptors) are extracted from both images at the detected keypoint locations and then matched exhaustively. The SIFT [13] algorithm is used as it has been proven to be robust and efficient. In order to further improve the matching, we use the recent RootSIFT variant proposed by Arandjelovic *et al.* in [14].

A homography matrix H is then estimated via the RANSAC [15] algorithm from the matched keypoints. To robustly compute the homography model parameters with RANSAC, the Symmetric Transfer Error (STE) [16] is used to measure the distances (the $l2$-norm here) between matched keypoints. Since the STE takes into account both the forward and backward projections of the matched keypoints, this distance is well suited for real-world data where local features detection and their matching will likely contain errors [16].

To increase the robustness of the estimation, the determinant of the homography matrix is also used to discard invalid models. As pointed out by Vincent *et al.* in [17], homographies with high or low determinants can be discarded as they correspond to degenerated cases, *i.e.* when the absolute value of the determinant of the matrix (or its inverse) is close to zero. Following the recommendation of [17], a threshold of 10 is used.

B. Multiple homography models estimation

A single homography is not always sufficient to describe the possible geometric distortions between frames, especially for non-planar scenes and scenes with strong parallax. To robustly estimate multiple homography models between a reference image and a target image, we propose to use a region-based estimation [12], consisting of a semi-local approach to better capture correlation between the two images. The main idea is to detect the major regions (planes or objects) found in both images and estimate the geometric transformations between them. We present here the main steps of the multiple models estimation, more details about the method can be found in [12].

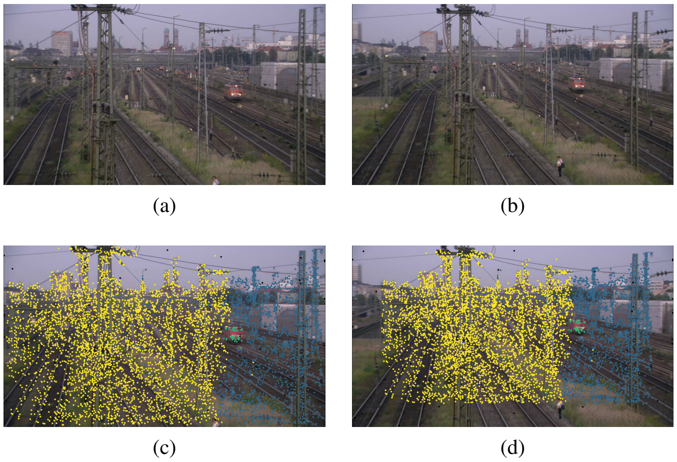


Fig. 1. Region-based estimation process: (a) current frame, (b) reference frame, (c) and (d) matched descriptors.

The target image is first segmented into small homogeneous regions by leveraging the efficient SLIC super-pixels algorithm [18]. SLIC uses a combined colorimetric and spatial distance metric to cluster similar and neighbouring pixels.

For each extracted super-pixel, a projective transformation, *i.e.* a homography model, is estimated from the SIFT keypoints located inside the super-pixel boundaries. We use the single homography estimation method previously presented in subsection II-A. However, some super-pixels may not have enough keypoints to estimate correctly a homography, may contain outliers, or may share similar or close homographies.

To improve the robustness as well as to reduce the number of homographies, the extracted models are recursively re-estimated and fitted to the matched keypoints via the energy minimization method proposed in [19]. The expansion (assignment) and re-estimation steps are performed iteratively until convergence of the minimization or until a maximum number of iterations is reached. At each iteration, the keypoints are assigned to the homography model that minimizes a combined energy function, the sum of three terms: data cost, smoothness cost and label cost. The data cost is a fidelity term, computed from the STE, which ensures that the model properly describes the projection. The smoothness cost is defined from the Delaunay triangulation of the matched keypoints. It penalizes neighboring points with different assigned homographies in order to preserve spatial coherence. The label cost is used to restrict the number of models. Due to the likely presence of outliers in the matches, an additional model is introduced to fit their distribution (a priori estimated). An example of the resulting labelling is shown in Figure 1d, where one can observe that 3 regions are detected successfully.

C. Photometric compensation

Once the finite set of homographies describing the geometric transformations between frames has been determined, predictions can be constructed. However, disparities due to illumination differences between the constructed blocks image

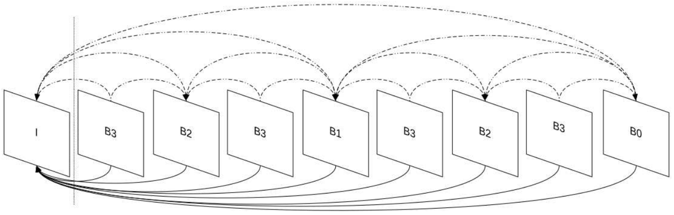


Fig. 2. Coding scheme: illustration for a GOP of 8 frames with the default random access configuration. The dashed lines represent the default references, the straight lines the extra reference for the proposed mode.

and the current block to encode may exist. During the encoding, these disparities will result in a highly energetic residue, limiting the use of the predictor by the encoder.

To compensate these distortions, we estimate an affine photometric compensation model for each previously estimated region. The two model coefficients, α and β are computed by minimizing the sum of absolute errors (SAD) on the matched pixel regions with a linear solver:

$$\underset{\alpha,\beta}{\operatorname{argmin}} \sum_{P} |Y'(x'_p) - (\alpha.Y(x_p) + \beta)| \tag{2}$$

This scale-offset model is estimated and applied only on the main component channel (Y) of the prediction blocks. We use the sum of absolute differences (SAD) to decide whether to enable or not the photometric model for each region during the prediction. If selected, the compensation is performed for all the blocks in the region. The SAD is used in this context as a fast estimator for the quality of reconstruction, compared to the full RDO process. As the SAD tends to favour more compact residues, it is preferred over the sum of squared differences (SSD).

III. Coding scheme

When compressing a video sequence with a classical video codec, most of the sequence frames are coded with the inter-prediction mode enabled, to make use of the temporal redundancy. All frames coded with inter-prediction leverage a set of reference frames, from which block predictions are performed by estimating and transmitting motion vectors and residues in the bit-stream. The proposed region-based estimator was implemented in this context.

In our setup, the region-based models are estimated between the original current frame and the original first frame of the group of picture (GOP). Although the predictor could be used for all the frames in the reference pictures buffer of the current image, we choose to use only one frame for implementation reasons, focusing on demonstrating the efficiency of the proposed mode. Potential prediction blocks are generated for each extracted model by warping and interpolating the reconstructed (encoded/decoded) blocks of the reference frame from the homography model, then compensating the luminance channel.

Once all the blocks of the current frame have been encoded, we determine which models have been actually used by the encoder through the rate distortion optimization (RDO) process. The default inter-prediction modes of HEVC are often efficient enough to predict the blocks and sending the parameters for multiple models is more expensive than simple motion vector parameters. As such, our mode competes with the other inter-prediction modes in the RDO loop and is only activated when the classical translational estimation fails to predict correctly the current block.

A specific syntax is added in the HEVC bit-stream to signal the used models, so that the stream can be decoded. The geometric and photometric models parameters are also encoded and stored in the bitstream for each frame, as half-precision floating point (16 bits).

TABLE I
BD-rate reduction compared with HEVC. Affine sequences in set (1), non-affine in set (2).

Sequence		**GLB**	**GLB-L**	**RB**	**RB-L**
(1)	CStoreGoods_720x1280	-2.29%	-2.29%	-2.58%	**-2.71%**
	DrivingRecorder1_720x960	0.16%	0.16%	-0.90%	**-1.05%**
	DrivingRecorder2_720x960	0.11%	0.07%	-1.63%	**-1.81%**
	LakeWalking_720x960	-7.06%	-7.06%	-9.26%	**-10.45%**
	ParkSunny_720x1280	-0.13%	-0.37%	-0.28%	**-0.74%**
	ParkWalking_720x1280	-2.04%	-2.04%	-2.73%	**-3.11%**
	bluesky_1920x1080	-3.90%	-4.21%	-3.97%	**-4.32%**
	station2_1920x1080	-11.28%	-11.75%	-12.13%	**-13.27%**
	tractor_1920x1080	0.53%	0.16%	**-0.27%**	**-0.27%**
	average	-2.88%	-3.04%	-3.75%	**-4.19%**
(2)	B_Cactus_1920x1080	0.02%	-0.03%	-3.16%	**-3.17%**
	city_1280x720	0.13%	0.13%	0.00%	**-0.26%**
	in_to_tree_1280x720	-0.28%	-0.28%	**-0.54%**	**-0.54%**
	shields_1280x720	-2.06%	-2.06%	**-2.13%**	**-2.13%**
	average	-0.55%	-0.56%	-1.46%	**-1.52%**
Total average		-2.16%	-2.27%	-3.04%	**-3.37%**

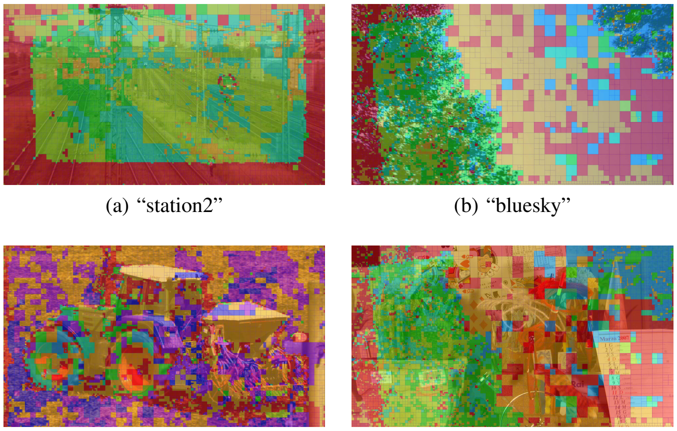
IV. Experimental results

The coding experiments are performed on common test sequences [9], [20] and proposed User Generated Content (UGC) sequences [21]. The selected sequences display a wide variety of motion caused by camera zooms, camera rotations, camera shakes, and classical 2D translational motion.

The HEVC HM software version 16.16 was used for all the experiments. The rate-distortion performances presented here are computed with the Bjontegaard metric [22] using the common 22, 27, 32, 37 Quantization Parameter (QP) values. The PSNR is computed on the Y channel only. The default HM *random access* configuration mode [9] is used as a baseline in all the following tests, with a GOP size of 16. The parameters for the region-based models estimator are fixed for all the experiments, more details about the adjustment of these parameters can be found in [12].

A. Coding results

Experimental results for the coding experiments are reported in Table I, the first set (1) of sequences corresponds to targeted sequences with known affine content, other sequences are placed in the second set (2). For comparison, we also introduce a global motion estimator as a second baseline, estimated


(a) “station2”

(b) “bluesky”

(c) “tractor”

(d) “Cactus”

Fig. 3. Region-based prediction mode usage. Red and orange blocks are coded with classical intra- and inter-prediction, respectively. Predictions made by our mode are depicted with the other colors.

with a classical SIFT+RANSAC approach (Section II-A). BD-rates reductions are presented for the global motion estimator (GLB), the global motion estimator with a global luminance compensation (GLB-L), the proposed region-based approach (RB) and the region-based approach with the luminance compensation (RB-L).

First, one can note that the GLB scheme brings an improvement of -2.88% over the default translational motion models of HEVC on the targeted sequences, highlighting the need for more complex motion prediction models. On the whole dataset, improvements go up to -11.75% with a mean BD-rate reduction of -2.16%.

The proposed region-based prediction mode achieves a greater improvement, with an average gain of -3.04%, up to -12.13%. Most of the sequences benefit from the multiple models prediction, with an average improvement of -0.88% over the single model mode. Although the gains are limited for most sequences, videos with affine motion display significant gains such as “station2” (-12.13%), “LakeWalking” (-9.26%) and “bluesky” (-4.32%).

The efficiency of the luminance compensation is low in the context of the global estimator, with only a marginal improvement of -0.11%, whereas a higher gain of -0.33% can be obtained with the region-based approach. As the luminance compensation is estimated from the matched regions content and not on the global frame, the estimation is more precise and robust.

Overall, our scheme achieves an average BD-rate gain of -3.37%, with -4.19% on targeted affine sequences and especially -1.52% on the second set.

To illustrate the use of the proposed models by the encoder, we adapted an HEVC bit-stream analyzer to display the use of the mode for each block. Examples are shown Figure 3 on 4 sequences. One may note that our prediction tool is enabled for a large number of the blocks within the reference frame “footprint”. For example, in the “station2” and “bluesky” sequences, the borders are not available for the prediction as a zoom and a rotation were respectively performed by the camera.

TABLE II
MEAN COMPLEXITY INCREASES COMPARISON AGAINST HEVC

Method	**Complexity**	
	Encoding	Decoding
GLB	189.92%	174.10%
GLB-L	190.27%	173.18%
RB	299.93%	196.01%
RB-L	301.45%	198.39%

B. Complexity study

We present here a brief complexity study of the proposed scheme. The prediction tool was implemented in the HM software (version 16.16) without particular optimization.

As it is often the case, the main complexity overhead of the prediction scheme resides on the encoder side. The mean complexity increases are reported in Table II. The mean runtime increase of the RB-L scheme is of ~300% compared to the default HM encoder. Most of the overhead is spend estimating the region-based models and in the increased RDO loop. Numerous improvements are possible to optimize the region-based estimator, especially the keypoints detection, extraction and matching process.

The complexity on the decoder side has a mean increase of ~200% for the RB-L method. Again, our implementation is not optimized. For example, we warp a whole frame for each model instead of warping only the selected blocks. Moreover, compared to the encoder, the decoder performs only a few extra operations. The model parameters are first decoded, then the blocks are generated by warping and interpolating the reference frame, and the luminance pixel values are finally corrected. All these operations can be optimized for fast processing. Besides, as the computation is still linear on the input ($O(n)$), an hardware implementation would have almost no overhead.

V. Conclusions

In this paper we present a novel video prediction mode to describe complex motion in video sequences. The efficiency of the proposed solution was demonstrated against state-of-the-art video coding tools on multiple sequences, with an average gain of -3.37% over HEVC. The complexity of the prediction mode is limited, especially on the decoder side, with respect to the gains that can be obtained. Although the region-based prediction is currently limited to one reference frame, it could be extended to use more frames from the reference pictures buffer. Improving the speed and the robustness of the prediction also constitutes an important future work.

References

[1] I. E. Richardson, *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*. John Wiley & Sons, 2004.

[2] F. Dufaux and J. Konrad, "Efficient, robust, and fast global motion estimation for video coding," *IEEE transactions on image processing*, vol. 9, no. 3, pp. 497–501, 2000.

[3] H. Watanabe and K. Jinzenji, "Sprite coding in object-based video coding standard: Mpeg-4," in *Proceedings of Multiconference on Systemics, Cybernetics and Informatics*, vol. 13. Citeseer, 2001, pp. 420–425.

[4] H. Jozawa, K. Kamikura, A. Sagata, H. Kotera, and H. Watanabe, "Two-stage motion compensation using adaptive global mc and local affine mc," *IEEE Transactions on Circuits and Systems for video technology*, vol. 7, no. 1, pp. 75–85, 1997.

[5] K. Rijkse, "H. 263: video coding for low-bit-rate communication," *IEEE Communications magazine*, vol. 34, no. 12, pp. 42–45, 1996.

[6] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 22, no. 12, pp. 1649–1668, 2012.

[7] H. Chen, F. Liang, and S. Lin, "Affine SKIP and MERGE modes for video coding," in *17th IEEE International Workshop on Multimedia Signal Processing, MMSP 2015, Xiamen, China, October 19-21, 2015*, 2015, pp. 1–5. [Online]. Available: https://doi.org/10.1109/MMSP.2015.7340829

[8] J. V. E. T. J. on Future Video Coding, "JVET JEM software," https://jvet.hhi.fraunhofer.de/svn/svn_HMJEMSoftware/.

[9] F. Bossen, "Common test conditions and software reference configuration," in *Proc. 12th JVT-VC Meeting*, Geneva, Switzerland, Jan, pp. 1–4.

[10] S. Parker, Y. Chen, D. Barker, P. de Rivaz, and D. Mukherjee, "Global and locally adaptive warped motion compensation in video compression," in *IEEE International Conference on Image Processing, ICIP 2017*, 2017.

[11] "Alliance for open media," http://aomedia.org.

[12] J. Bégaint, D. Thoreau, P. Guillotel, and C. Guillemot, "Region-based prediction for image compression in the cloud," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1835–1846, April 2018.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition, 2012*, 2012, pp. 2911–2918.

[15] M. A. Fischler and R. 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, 1981.

[16] A. Harltey and A. Zisserman, *Multiple view geometry in computer vision (2. ed.)*. Cambridge University Press, 2006.

[17] E. Vincent and R. Laganiere, "Detecting planar homographies in an image pair," in *ISPA 2001. Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis. In conjunction with 23rd International Conference on Information Technology Interfaces*, 2001, pp. 182–187.

[18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[19] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov, "Fast approximate energy minimization with label costs," *International Journal of Computer Vision*, vol. 96, no. 1, pp. 1–27, 2012.

[20] "Xiph.org test media," http://media.xiph.org/video/derf/.

[21] X. Ma, H. Zhang, Y. Zhao, M. Sun, M. Sychev, H. Yang, and J. Zhou, "Huawei test sequences of UGC feature for video coding development," in *Proc. 22th JVT-VC Meeting*, Geneva, Switzerland, Oct.

[22] G. Bjontegaard, "Calculation of average psnr differences between rd-curves," in *ITU-T SG16/Q6 VCEG document VCEG-M33*, Austin, TX, USA, Apr 2001.