Region-based prediction for image compression in the cloud

Jean Bégaint, Dominique Thoreau, Philippe Guillotel, Christine Guillemot

Region-based prediction for image compression in the cloud

Jean Bégaint, Dominique Thoreau, Philippe Guillotel and Christine Guillemot

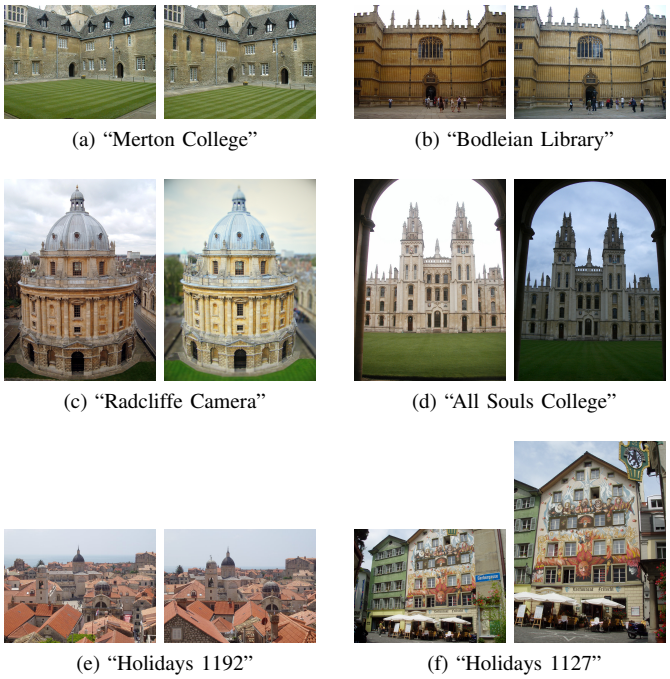
***Abstract*—Thanks to the increasing number of images stored in the cloud, external image similarities can be leveraged to efficiently compress images by exploiting inter-images correlations. In this paper, we propose a novel image prediction scheme for cloud storage. Unlike current state-of-the-art methods, we use a semi-local approach to exploit inter-image correlation. The reference image is first segmented into multiple planar regions determined from matched local features and super-pixels. The geometric and photometric disparities between the matched regions of the reference image and the current image are then compensated. Finally, multiple references are generated from the estimated compensation models and organized in a pseudo-sequence to differentially encode the input image using classical video coding tools. Experimental results demonstrate that the proposed approach yields significant rate-distortion performance improvements compared to current image inter-coding solutions such as HEVC.**

I. Introduction

The emergence of cloud applications and web services has led to an increasing use of online resources. Associated with the large availability of high-end digital cameras in smartphones, as well as the rise of online storage solutions (*e.g.* Google Photos, Flickr, OneDrive, Dropbox) and new social media practices (*e.g.* Facebook, Twitter, Pinterest), images and videos constitute today a significant part of this data. Billions of images are already stored in the cloud, and hundreds of millions are uploaded every day [1]. Furthermore, these images are rarely deleted and often duplicated across filesystems and datacenters to mitigate data loss risks.

Images are usually independently encoded with the classical JPEG [2] codec. However, given the amount of data saved in the cloud, very similar content may already be stored online and this redundancy can be exploited to significantly reduce storage requirements. Given a large enough dataset of images, a new image could then be encoded from a reference, or multiple references, already present in the cloud. An example of similar images that could be found in such a database is shown Fig. 1.

Inter-coding of images is traditionally used in video compression, where the redundancy is reduced by encoding consecutive frames from previous frames used as references. Solutions have been proposed to leverage the inter-prediction tools of video codecs to encode similar images as pseudo-video sequences [3], [4]. Still, video codecs are primarily designed assuming that rigid, block-based, two-dimensional displacements are suitable models for the motion taking place in a scene. In the considered case, disparities between correlated images can result from pictures taken from different viewpoints, with different cameras, focal lengths, illumination



(a) "Merton College" (b) "Bodleian Library" (c) "Radcliffe Camera" (d) "All Souls College" (e) "Holidays 1192" (f) "Holidays 1127"

Fig. 1: Example of targeted images presenting geometric distortions and illumination disparities.

conditions, at different points in time, etc. These disparities can be characterized by geometric transformations (*e.g.* translations, rotations, zoom) or photometric transformations (*e.g.* illumination disparities, gamma changes). Besides, image scenes are not always planar, and as such, multiple distortions can occur within an image pair. Several approaches have been successfully proposed to encode correlated images by compensating these distortions with multiple transformation models [5], [6]. However, in the work of Shi *et al.* [5], the number of transformation models is restricted to 4, whereas in the method of Zhang *et al.* [6] the frame is divided into 256×256 pixel prediction units. Furthermore, both of these methods do not take into account the image content. Thus, we introduce here a compression scheme able to efficiently handle non-planar images with complex deformations via a region-based approach.

In this paper, we propose a novel region-based prediction scheme able to leverage correlation between similar non-planar images. Unlike existing approaches, our scheme extracts multiple regions, planes or objects, of the current image, each subject to a distinct transformation model. Geometric and

photometric disparities are then efficiently compensated in a region-wise manner to predict the targeted frame, which is then finally encoded with HEVC [7]. As an alternative for a classical scale-offset compensation on the luminance channel, we also propose to apply a compensation model on the color channels, which is able to address larger disparities.

Experimental results indicate that the proposed scheme can efficiently leverage inter-image redundancy, achieving on average a 19.6% BD-rate reduction compared to HEVC inter coding, computed on a dataset of several hundreds sequences. We also demonstrate that our scheme is competitive in terms of bit-rate distortion performances when compared against state of the art methods.

The rest of this paper is organized as follows. Related work is reviewed in Section II. Section III gives an overview of the proposed compression scheme. Section IV describes it in details. Experiments are reported and discussed in Section V. Section VI concludes this paper.

II. Related work

A. Image set compression

Zou *et al.* propose in [3] to organize images from an album into a tree structure, then encode it as a video sequence. A graph structure is first determined by order of similarity between the images. Correlations are measured with the sum of square differences (SSD). An image tree is then obtained from the graph via a minimum spanning tree approach (MST) and encoded with HEVC, with a group of pictures (GOP) of one I frame (the tree root) and n following P frames, *i.e.* the leaves. The scanning is performed via a depth-first search algorithm, meaning the lowest leaves are explored before going upwards. A maximum depth of the tree is imposed in order to limit the image retrieval time (the random access). The authors obtain an overall improvement of 75% over JPEG. However, this method relies on the SSD for measuring the correlation, which is not robust to geometric and illumination changes. In addition, accessing a random image requires prior decoding of several images and increases the loading time. Moreover, video encoders have not been designed to cope with variations in terms of focal length, viewpoint, illumination, encountered in sets of images.

When considering millions of images available in the cloud, it is very likely that from a given image, another highly similar image can be found in a very large database [4]. Perra *et al.* thus propose to take advantage of the large online datasets and the inter-coding performance of HEVC to compress image pairs, and introduce a novel approach with low computational cost. To find correlated images, global feature descriptors are used. A GIST descriptor [8] is computed from the current image and then reduced to a 512 bits representation. GIST descriptors have been selected as they are as efficient as SIFT in this context [9], and with a lower computational cost. A nearest neighbour search (K-NN) is then performed to retrieve the most correlated image from the dataset. An HEVC inter-coding is finally applied with the reference image as an I frame and the query image as a P frame. This method provides fast operations, suitable for online applications, and produces an average reduction of size by a 74% factor with a canonical set of 13 million images, compared to JPEG.

B. Feature-based image compression

Additional methods have been developed to deal with sets of images with larger disparities. As such, Yue *et al.* propose in [10] to encode an image from its down-sampled version and local feature descriptors. The descriptors are used to retrieve correlated images from the cloud and identify corresponding patches. As an image can have thousands of SIFT descriptors [11], the total size of feature vectors can exceed the image size. The SIFT descriptors of the current image are thus encoded from the SIFT descriptors extracted from the down-sample version of the image. Only the compressed descriptors and the encoded down-sampled image are then sent to the cloud. Once the data has been decompressed, the image can be reconstructed. First, highly correlated patches are retrieved from the cloud. The transformation between a pair of patches (retrieved and up-sampled) is estimated by applying the RANSAC algorithm on the descriptors. Finally, the patch stitching is guided by the up-sampled decompressed image. This method achieves an average 1885:1 compression rate, and yields a better subjective quality than JPEG and HEVC intra-coding. However, this method has some limitations. On some images, one may not find sufficiently correlated images in the cloud. Complex images can also be too difficult to reconstruct faithfully. The authors propose then to extract the complex parts of the image and encode them with classical image compression tools. Furthermore, this method requires high computational power to perform all the operations. Although good visual results and an impressive compression ratio can be obtained, this technique might not reconstruct faithfully the original image due to the use of sparse local feature, and the absence of residual coding.

Another approach has then been proposed by Shi *et al.* in [5], relying on local feature descriptors. They introduced a three-step method to reduce inter-image redundancy. A feature-based multi-model approach is first used to compensate geometric transformations between images. Then, a photometric transformation is applied to account for illumination changes between the references and the target image. Finally, a block matching compensation (BMC) is performed to compensate remaining local disparities. To evaluate the geometric transformation, a content-based feature matching is first performed by using SIFT local feature descriptors [11]. The matching between images is performed based on the correlation between groups of descriptors instead of pixel values. A K-means algorithm is applied to cluster SIFT descriptors and organize the images into correlated sets. The images are placed in a graph, the weights are computed as the distance between matched SIFT feature vectors. The prediction structure is obtained by converting the graph into a MST. The number of transformations and their parameters are then derived, and the geometric transformation is then estimated via the RANSAC algorithm. A feature-based photometric compensation is proposed to compensate illumination changes. Finally, a BMC is used to account for local disparities, which

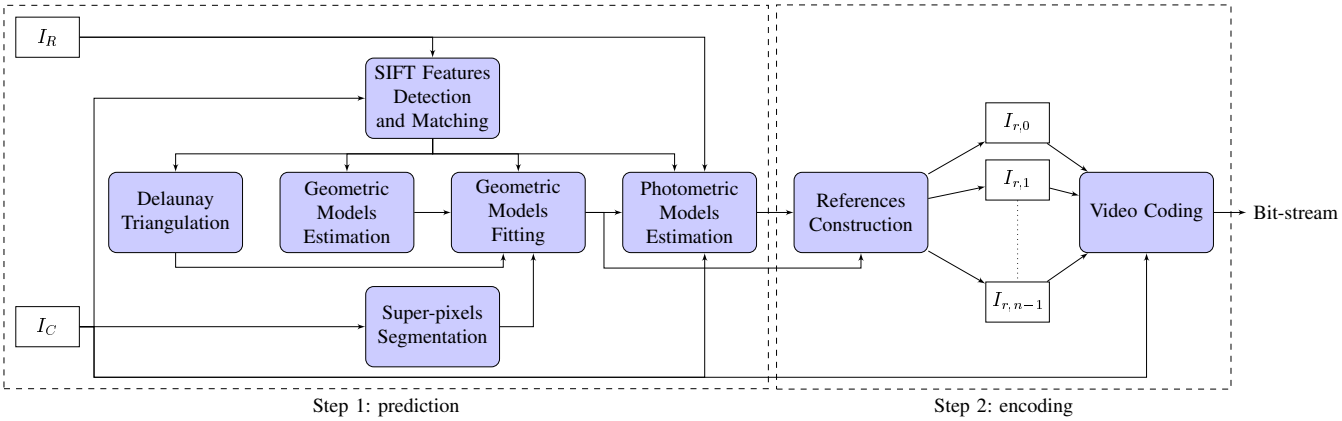


Fig. 2: Illustration of the proposed compression scheme.

are not compensated by the global geometric and photometric compensations. This method outperforms JPEG by reducing the bit-stream size by a factor of 10, while maintaining the same quality.

It is worth also pointing out that similar registration techniques based on local features extracted from correlated images have also been used successfully for image super-resolution [12]–[17] and image denoising [18], [19] tasks.

Recently, Zhang *et al.* presented in [6] a novel prediction method based on dense correspondences. They proposed to compensate geometric and photometric distortions on a 256×256 pixels block basis. By using dense pixel to pixel correspondences in local units instead of local descriptors, the parametric estimation of the geometric models and the luminance compensation is more robust to local disparities.

In a previous work [20], we proposed a method relying on a global compensation associated to a local prediction based on locally-weighted template matching. Compared to current coding solutions, significant rate-distortion performance improvements have been obtained, at the cost of high complexity.

In this paper, we present a different approach based on a semi-local prediction model which relies on a region-based estimation of multiple homographic and photometric models.

III. Overview of the proposed compression scheme

The proposed compression scheme comprises two main steps, as shown in Fig. 2. For the purpose of explanation, we will only consider a pair of images but our scheme can also be adapted for larger sets of images. When considering the current image I_C to be encoded, a reference image I_R is first retrieved from the cloud with the help of a classical Content Based Image Retrieval (CBIR) system. Additional reference images $I_{r,i}$ are then constructed by exploiting geometric and photometric transformation models between the reference and the current images. The current image I_C is finally encoded from the reference images $I_{r,i}$ with a video encoder such as HEVC. To decode I_C, the reference images $I_{r,i}$ are reconstructed from the reference image I_R and the transformation models. The reference image thus needs to be available both at the encoder and the decoder sides. In this paper, we assume that the reference image is retrieved from a large and static image database, and is referenced in the bit-stream.

The proposed prediction method relies on a semi-local approach which estimates region-based geometric and photometric models to better capture correlation between the two images. To segment the current image into homogeneous regions, in terms of geometric transformations, the image is first segmented into super-pixels. SIFT descriptors are then extracted from both images and matched exhaustively. For each super-pixel extracted from I_C, a projective transformation, *i.e.* a homography model, is estimated from the SIFT keypoints located inside the super-pixel boundaries. To reduce the number of homographies the estimated models are recursively re-estimated and fitted to the keypoints via the energy minimization method proposed in [21]. The Delaunay triangulation of the keypoints is used to preserve the spatial coherence during the homographies estimation. Then, the photometric disparities between I_C and I_R are compensated region-wise by estimating a transformation model between matched regions of the image pair. Multiple references $I_{r,i}$ are generated by warping each region using its assigned homographic model and applying the photometric compensation. Finally, the references are organized in a pseudo-sequence in which the current image is differentially-encoded with classical video coding tools. The side information (SI), *i.e.* the homographies and the photometric model coefficients required to reconstruct the predictions on the decoder side, need to be transmitted and are taken into account in the bit-rate.

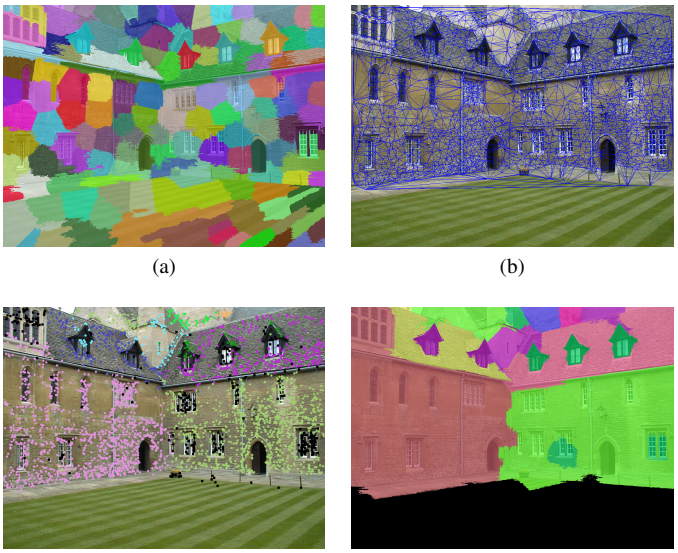
IV. Region-based prediction scheme

A. Super-pixel segmentation

To initialize the region-based segmentation, a super-pixel segmentation is first performed via the SLIC algorithm proposed by Achanta *et al.* in [22]. All the pixels i of I_C are clustered according to a combined colorimetric and spatial distance $D(C_k, i)$ to a centroid C_k defined as

$$D(C_k, i) = \sqrt{\left(\frac{d_c}{m_c}\right)^2 + \left(\frac{d_s}{m_s}\right)^2} \quad (1)$$

where d_c represents the $l2$-norm in the LAB colorspace, d_s the l_2 norm between a given pixel i and a centroid



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

Fig. 3: Region-based geometric estimation: (a) SLIC segmentation of I_C. (b) Mesh of the Delaunay triangulation of the matched keypoints. (c) Keypoints labels, each keypoint is assigned a homography model, the outliers points are represented in black. (d) Region segmentation.

C_k. The quantities m_s and m_c are weighting parameters used to normalize color and spatial proximity. Our scheme relies on the Adaptive-SLIC (ASLIC) variant of the SLIC algorithm, where m_s and m_c are updated at each iteration of the algorithm. When using SLIC, m_s and m_c are set to constant values, the assumed maximum colorimetric and spatial distance. Whereas with ASLIC, only the first iteration relies on fixed normalization parameters, they are then updated to the maximum distances observed in each cluster at the previous iteration. According to [22], this decreases the boundary-recall performance. However, the super-pixels compactness parameter is highly dependent of the image content and its contrast. Thus, by using the adaptive version of the algorithm, no per-image tuning is required, since the initial parameters are updated along the iterations.

The SLIC segmentation is initialized from a regular grid of centroids $\{C_k | k \in [0, K[\}$ spaced by a fixed distance, the step size s, and result in a segmentation of n super-pixels. With $K = \lfloor \frac{w}{s} \rfloor * \lfloor \frac{h}{s} \rfloor$, (w, h) the image size, and $n \leq K$ depending on the clean-up step, where some centroids with too few assignments can be removed.

An example of the resulting segmentation of the current image I_C is shown in Fig. 3a.

B. Geometric model estimation

To estimate the geometric models, our scheme relies on local feature descriptors as they are more robust to geometric distortion (*e.g.* translation, rotation, zoom, scale) and illumination variations than the pixel values [11].

SIFT keypoints are first extracted from both I_C and I_R and then matched exhaustively. In order to improve the matching, we use the RootSIFT algorithm proposed by Arandjelovic *et al.* in [23]. The computed SIFT descriptors X_i are first projected into a feature space:

$$X_i^{'} = \sqrt{\frac{X_i}{\|X_i\|_1}}, \forall i \in [\![1, N]\!] \quad (2)$$
$$with \ \ \|X_i\|_1 = \sum_{j=1}^{128} |X_i(j)|$$

then the distance between them is computed using the l_2 norm. For each super-pixel, a homography model H, defined by the matrix

$$H = \begin{bmatrix} s_x.\cos(\theta) & -s_y.\sin(\theta+\sigma) & t_x \\ s_x.\sin(\theta) & s_y.\cos(\theta+\sigma) & t_y \\ k_x & k_y & 1 \end{bmatrix} \quad (3)$$

is then estimated via the RANSAC [24] algorithm from the matched keypoints contained within the super-pixel boundaries. Here (t_x, t_y) denote 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.

RANSAC is an iterative method which can estimate a parametric model from a noisy set of data points. There is no guarantee that the optimal solution will be found during the iterations. However, the probability of success is independent of the number of points in the data set and only relies on two parameters: the number of iterations N and the residual threshold t to discard an outlier. Let u be the probability of a data point to be an outlier, the minimal number of iterations to reach a probability p of finding the optimal solution is given by

$$N = \frac{\log(1-p)}{\log(1-u^c)} \quad (4)$$

where c is the minimum number of samples to estimate the parametric model. In the case of a homography model, $c = 4$ (8 degrees of freedom).

To robustly estimate a homography model with RANSAC, the Symmetric Transfer Error (STE) [25] is used to compute the distances between matched keypoints:

$$STE(H_l) = \overbrace{\sum_{p \in P} d(x'_p, H_l.x_p)^2}^{\text{forward term}} + \overbrace{\sum_{p \in P} d(x_p, H_l^{-1}.x'_p)^2}^{\text{backward term}} \quad (5)$$

where H_l denotes a homography model to be evaluated, x_p and x'_p two matched keypoints, and d the euclidean distance. Since the STE takes into account both forward and backward projections of matched keypoints, this distance is well suited for real-world data where local feature detection and their matching will likely contain errors [25].

To further improve the estimation process, the determinant of the homography matrix is also used to discard invalid models. As pointed out by Vincent *et al.* in [26], homographies not respecting the condition:

$$\mathcal{H} = \left\{ H_l \mid \frac{1}{k} \leq |\det(H_l)| \leq k \right\} \tag{6}$$

can be rejected as they correspond to degenerated cases, *i.e.* the absolute value of the determinant of the matrix (or its inverse) is close to zero. Following the recommendation of [26], we set k to 10.

From the n super-pixels of the SLIC segmentation, m homography models are thus estimated, with $m \leq n$. Indeed, some super-pixels do not contain a sufficient number of matched keypoints to estimate a projective transform, or contain only outliers. Furthermore, the models attributed to neighboring super-pixels may be very similar as they might be part of the same region.

C. Geometric model fitting

From the previously estimated homography models, the most representative model for each region needs to be extracted and refined before generating the projections.

Delong *et al.* proposed in [21] an efficient method to solve the issue of multiple models fitting. To solve this labelling problem, *i.e.* assigning a model to each keypoint, they introduce a new joint discrete energy:

$$E(f) = \overbrace{\sum_{p \in P} D_p(f_p)}^{\text{data cost}} + \overbrace{\sum_{(p,q) \in N} V_{pq}(f_p, f_q)}^{\text{smooth cost}} + \overbrace{\sum_{L \subseteq \mathcal{L}} h_L . \delta_L(f)}^{\text{label cost}} \tag{7}$$

to be minimized iteratively, where N is the keypoints neighborhood, h_L the label cost of the subset of labels L, and where the function $\delta_L(f)$ is defined as:

$$\delta_L(f) \triangleq \begin{cases} 1, & \exists p : f_p \in L \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Following the set-up described in [27], an initial proposal for the homography models needs to be estimated from the matched keypoints, the observations P. During the expansion step, each keypoint p is assigned a label l from the set of homographies L in order to minimize the objective function (7). From the labelling f, the set of models can then be updated (re-estimation step). The expansion and re-estimation steps are performed iteratively until convergence of the minimization of (7) or until a maximum number of iterations is reached.

In the set-up described in [21] and [27], the set of initial homography models is randomly generated by selecting N samples of 4 matches. In our approach, we use the models previously estimated from the super-pixels, which allows for a faster convergence and a more robust estimation. The set of homography models is then reduced and refined by recursively minimizing the energy (7).

The data cost is a fidelity term, which ensures that the model properly describes a transformation, computed from the STE (5). Due to the likely presence of outliers in the matches, an additional model ϕ is introduced to fit their distribution, with a fixed data cost for all the vertices and a label cost set to zero:

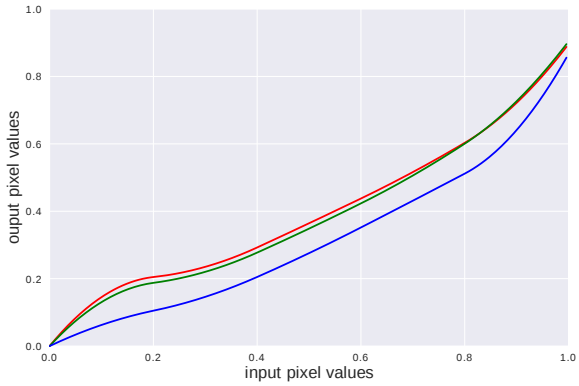


Fig. 4: Example of the splines fitting on the RGB channels.

$$\begin{cases} h_\phi = 0 \\ D_p(\phi) = C, \text{with } C > 0 \end{cases} \tag{9}$$

The smoothness cost for the set of neighbours $pq \in N$ is defined from the Delaunay triangulation of the matched keypoints in the current image I_C (Fig. 3b). It penalizes neighboring points with different labels in order to preserve spatial coherence and is defined as:

$$V_{pq} = w_{pq} \cdot \delta(f_p \neq f_q) \text{ with } \begin{cases} w_{pq}, & \text{weight for the vertex } pq \\ \delta, & \text{Kronecker delta} \end{cases} \tag{10}$$

The label cost (8) is used to restrict the number of models.

An example of the resulting labelling is shown in Fig. 3c, where one can observe that several planes, or regions, of the image are detected successfully.

D. Photometric compensation

Once the finite set of homographies describing geometric transformations between image pairs has been determined, a reference image can be constructed. However, disparities due to illumination and photometric differences between the constructed reference image and the current image persist. During the encoding, these disparities will result in a highly energetic residual, limiting the use of the predicted image by the encoder.

To compensate these distortions, we propose to estimate a photometric compensation model for each previously estimated region.

A scale-offset model is often proposed to minimize distortion on the Y channel ([5], [6], [10]). The model coefficients, α and β are computed by minimizing the sum of square errors on the matched keypoint pixels:

$$\underset{\alpha,\beta}{\operatorname{argmin}} \sum_{P} |Y'(x'_p) - (\alpha . Y(x_p) + \beta)|^2 \tag{11}$$

This model can efficiently handle illumination disparities, but performs poorly on complex colorimetric disparities. We

choose to add the more flexible model proposed by Hacohen *et al.* in [28]. The photometric deformation is modelled by a piece-wise cubic spline f on each RGB channel. This model can compensate for a variety of photometric distortions such as gamma changes or color temperature. The minimization problem:

$$\begin{aligned} \underset{f}{\operatorname{argmin}} \quad & \sum_{Q} |I'(x'_q) - f(I(x_q))|^2 + C_{soft}(f) \\ \text{subject to:} \quad & C_{hard}(f) \end{aligned} \tag{12}$$

is solved for 6 knots (0, 0.2, 0.4, 0.6, 0.8, 1) via quadratic programming. The same soft constraints (C_{soft}) and hard inequalities constraints (C_{hard}):

$$\begin{aligned} C_{soft}(f) = \lambda_1 & \sum_{x \in \{0,1\}} |f(x) - x|^2 \\ + \lambda_2 & \sum_{x \in \{0.2j-0.1\}_{j=1}^{5}} |f(x) - x|^2 \\ + \lambda_3 & \sum_{x \in \{0.2j-0.1\}_{j=1}^{5}} |f''(x)|^2 \end{aligned} \tag{13}$$

$$C_{hard}(f) = \begin{cases} 0.2 \leq f'(x) \leq 5, \forall x \in \{0.2j - 0.1\}_{j=1}^{5} \\ f(0) \leq 0 \end{cases} \tag{14}$$

are used to control smoothness and monotonicity of the curves. Hard equality constraints are also set on the 4 inner knots of the splines and their first derivative. Each curve thus has 7 degrees of freedom.

The minimization is performed for each region determined from the labelling. As we cannot rely on a dense correspondence field as in the original paper [28], we use a set of pixels Q within a given radius of matched keypoints of each region, to ensure that only reliable pairs of pixels values are used.

We use the sum of absolute differences (SAD) to select the best performing photometric model for each region during the prediction. The SAD is preferred here over the sum of squared differences (SSD), as it tends to favour more compact residuals, and thus is considered as a better estimator for the quality of the reconstruction. The photometric compensation can also be disabled when the image pair does not present any photometric distortions or the estimation fails.

E. Pseudo-video sequence encoding

Once the geometric and photometric models have been successfully estimated, the image can be segmented into regions at the pixel level. The region segmentation is computed by selecting the best projection for each super-pixel. The mean absolute error is used to measure the distortion for each super-pixel between a given projection and the current image. An example of the final segmentation is shown in Fig. 3d.

A prediction image can then be constructed from the reference frame, the estimated models and the region segmentation. However, sending this segmentation map to reconstruct the prediction on the decoder side would be costly. Instead, multiple reference pictures are used, which can be constructed in the same manner on both the encoder and decoder sides.

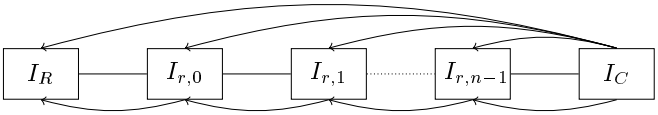


Fig. 5: Illustration of the pseudo-video sequence encoding scheme used.

TABLE I: Side information (SI) sent to the decoder for each of the n predicted regions. For the photometric compensation, one of the two models is chosen for each region, or the compensation is disabled.

Compensation method	**Model**	**Bits**
Geometric	*homography*	$8 * 16$
Photometric	*scale-offset*	$2 * 16$
	piece-wise spline	$7 * 16$

Those n additional references $I_{r,i}$ are constructed from the reference image I_R and the region models (the associated geometric and photometric compensation models). For each region, an additional reference image $I_{r,i}$ is constructed by warping the reference image I_R with the associated region homography model and applying the photometric correction. This step is both performed at the encoder and decoder side, and as such the encoded $I_{r,i}$ are discarded in the transmitted bit-stream. The reference image, the projections and the current image are then concatenated in a pseudo-video sequence, finally encoded with HEVC. Our encoding structure differs from the main HEVC profiles such as the low-delay and hierarchical configurations, as the last frame needs to be predicted from all the previous frames in the sequence in order to fully exploit the inter-redundancies.

Starting from the low-delay configuration of the HM software (*lowdelay_P_main.cfg*), the GOP settings are modified to keep all the frames in the reference pictures buffer, as shown in Fig. 5. The reference frames are encoded at maximum quality ($QP = 0$), since this part of the bit-stream will not be stored in the final bit-stream, while the quality of the current frame is controlled via the $QPoffset$ value.

To enable the decoder to reconstruct the projections used as reference pictures for the current image, some Side Information (SI) is also stored alongside the HEVC bit-stream. By using multiple reference frames, only the geometric and photometric models coefficients need to be transmitted. The encoder then performs its reference selection for each inter-coded prediction unit and stores it in the bit-stream. This avoids sending the costly segmentation map, and lets the encoder decide the best reference frame to select for each prediction-unit, in the rate distortion optimization (RDO) loop. All the SI parameters are stored as half-precision floating point, coded on 16 bits each. For the homography models, 8 parameters need to be stored in the bit-stream, 2 parameters for the scale-offset model or 7 parameters for each color channel for the piece-wise spline fitting model, as detailed in Tab. I.

V. Experimental results

To perform our experiments, numerous images have been retrieved from publicly available databases [29]–[32] and also crawled from Google Images and Flickr. The collected images present challenging disparities such as combinations of different viewpoints, focal lengths, illumination variations, translations, rotations. Such disparities result from pictures taken at different points in time, with different camera positions, lighting conditions, etc...

Unless otherwise specified, the HEVC HM[1] software version 16.9 with the *low-delay* configuration is used for the video coding in all the following tests. The rate-distortion performances presented in the rest of this paper are computed with the Bjontegaard metric [33] using the recommended settings of 22, 27, 32 and 37 for the Quantization Parameter (QP). The PSNR is computed on the Y channel.

The super-pixel centroids are initialized on a regular grid, spaced by 64 pixels. The initial compactness is set to 10. In order to use the energy (7) to estimate the multiple geometric models, the value of the label, smooth and outlier costs first need to be determined. As stated by Delong *et al.* in [21], these parameters are application dependent, and as such, they can be learned offline once, on a representative dataset. Their values have been computed on a training dataset with the differential evolution algorithm introduced by Storn & Price in [34]. This method allows finding the global minimum of a multivariate function, over a large space of possible parameters combinations more robustly than with manual tuning, to the detriment of a slow convergence.

As regards the splines fitting based photometric model, the parameters provided in [28] have been used. The pixel search radius is set to 15 pixels and the quadratic problem has been solved with an efficient quadratic solver[2].

The same set of parameters is used for all the results presented in this paper.

A. Performance of region-based models

The performance of the region-based prediction model ("region-based") is first compared with the performance of a global compensation model ("global") and also with HEVC low-delay ("inter"). The two prediction modes are evaluated with only the geometric compensation enabled ("geo") and both the geometric and photometric compensations enabled ("geo+photo"). The global compensation scheme consists in a single homography transformation, estimated from the classical SIFT+RANSAC approach, followed by a photometric scale-offset compensation, *i.e.* with only one homography and one photometric compensation model per image. Examples are shown in Fig. 6 for the four sequences of Fig. 1: "Merton College" (Fig. 6a), "Bodleian Library" (Fig. 6b), "Radcliffe Camera" (Fig. 6c), and "All Souls College" (Fig. 6d), the Bjontegaard metrics are provided in Tab. II. The "Merton College" sequence exhibits multiple geometric distortions, while "Bodleian Library" and "Radcliffe Camera" both present multiple geometric and photometric distortions, and "All Souls College" a global geometric and photometric distortion. The choice of which image is the current/reference image was performed randomly.

The proposed scheme results in the following respective BD-rate improvements of 28.50%, 39.80%, 61.32% and 40.54% compared with HEVC inter. For the "Merton College" sequence, the BD-rate gain increases from 9.96% to 28.50% thanks to the use of multiple geometric compensation models, whereas the photometric compensation does not yield any performance improvements for this sequence.

On the "All Souls College" sequence, one might observe that the photometric compensation can greatly improve the efficiency, from 10.63% to 38.63%. Also, in this case, the photometric compensation of the region-based model is more performant, from 38.63% to 61.32%. Although there is only one region in the image, the photometric model based on splines yields a better prediction. This is confirmed on the "Bodleian Library" and "Radcliffe Camera", which all benefit from the photometric compensation, from 28.03% to to 39.80% and from 26.03% to 40.54%, respectively. It is also worth noting that on the "Radcliffe Camera" and "Bodleian Library" sequences, the global scheme with the photometric compensation performs better than the region-based algorithm without photometric compensation, emphasizing its crucial role in providing an accurate prediction.

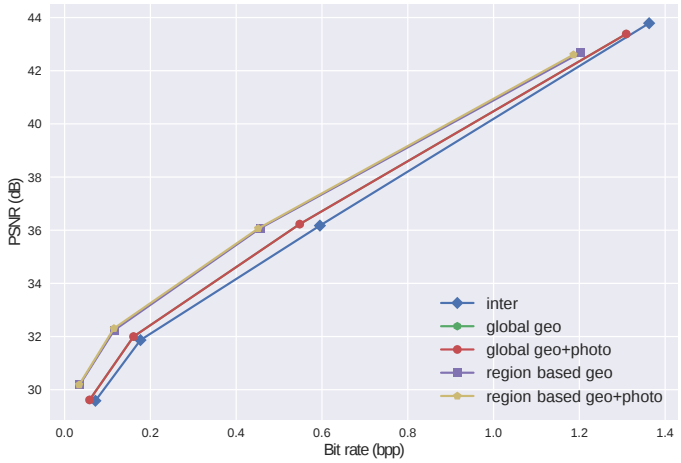
For the four sequences, the respective bit-stream ratio allocated for the side information over the total bit-stream size is 0.77%, 0.41%, 0.28% and 0.25%, which is negligible.

Results for the reversed sequences, where the reference and current images are swapped, are also provided for comparison. The rate-distortion improvements are consistent for the "Merton College" and "Radcliffe Camera" sequences. However, while an increase in performance is observed for the "Bodleian Library" sequence, one can notice a decrease for the "All Souls College". This can be explained by the different exposures of the frames in each sequence. Indeed, predicting an image from an under-exposed correlated image is more challenging, as numerous details are lost due to the lack of brightness and thus cannot be predicted correctly.

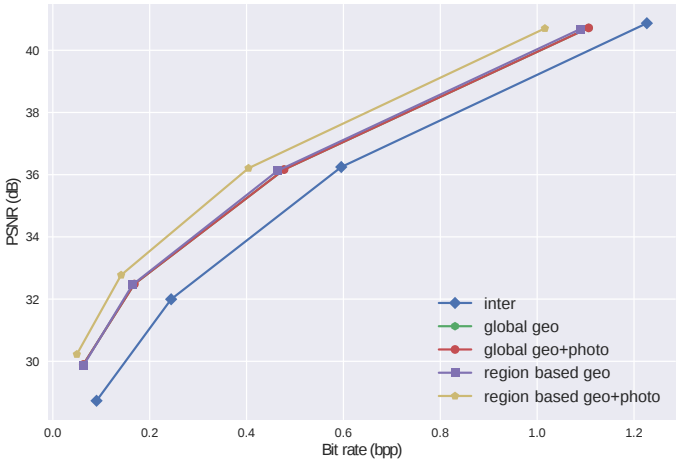
To illustrate the performance gains due to the use of the predicted regions, we modified an HEVC bit-stream analyzer to display the reference picture index used for each coding unit. An example is shown Fig. 7 on the "Merton College" sequence, where the encoder decisions for the reference picture selection are displayed for different QP values. Each color indicates a reference frame, *i.e.* a region, chosen by the encoder as a reference picture, the intra mode is represented in black. One can observe that the reference frame selection for each coding unit in the quad-tree is overall quite consistent with the region-based segmentation presented previously. Still, there are some local inconsistencies in the reference selection that can be attributed to the decisions of the RDO loop. Also, at higher bit-rates, *i.e.* lower QP values, the intra-mode is more frequently selected by the encoder, especially in complex zones such as the windows where the light reflection cannot be predicted accurately.

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

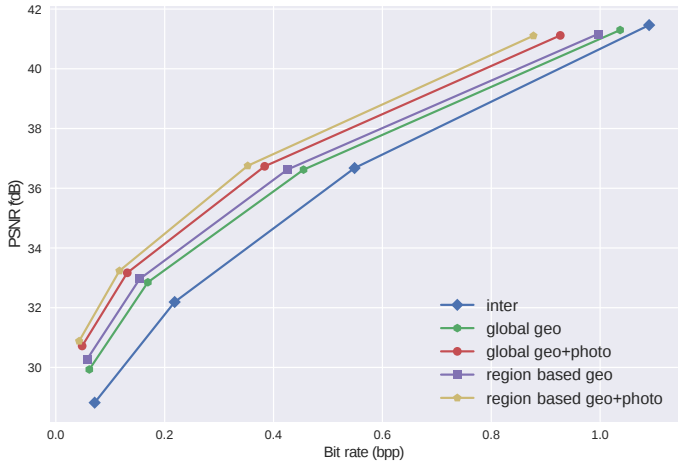
[2] https://github.com/liuq/QuadProgpp



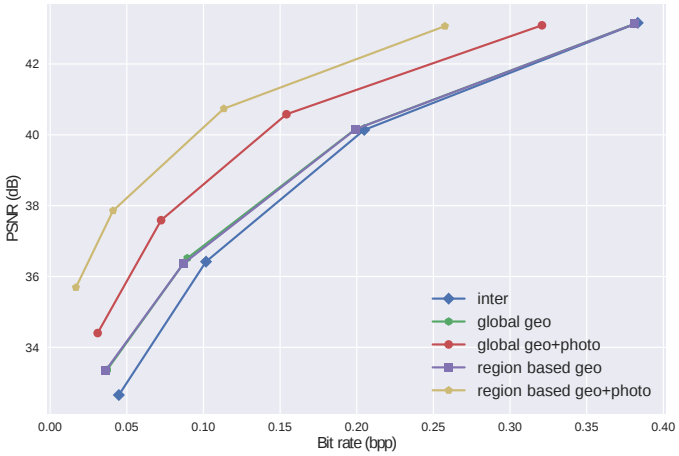
(a) "Merton College"



(b) "Bodleian Library"



(c) "Radcliffe Camera"

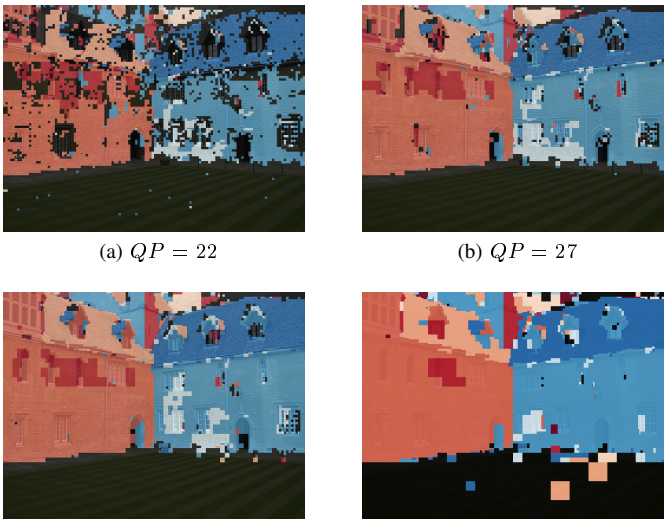


(d) "All Souls College"

Fig. 6: Performance comparison of the different prediction methods. The proposed region-based model and a global scheme are compared to HEVC inter, with and without photometric compensation.

TABLE II: Bjontegaard metrics computed on the rate-distortion curves presented in Fig. 6. The symbol "*r*" indicates that the sequence was processed backward, *i.e.* the reference image and the current image were switched.

Sequence	**global geo**		**global geo+photo**		**region-based geo**		**region-based geo+photo**	
	BD-rate	BD-PSNR	BD-rate	BD-PSNR	BD-rate	BD-PSNR	BD-rate	BD-PSNR
"Merton College"	9.96%	0.42db	9.96%	0.42db	**28.50%**	**1.22db**	**28.50%**	**1.22db**
"Bodleian Library"	26.94%	1.40db	26.94%	1.40db	28.03%	1.47db	**39.80%**	**2.17db**
"Radcliffe Camera"	17.78%	0.90db	34.18%	1.84db	26.03%	1.34db	**40.54%**	**2.26db**
"All Souls College"	10.63%	0.51db	38.63%	2.18db	10.55%	0.50db	**61.32%**	**3.79db**
"Merton College" *r*	15.52%	0.74db	15.52%	0.74db	**28.32%**	**1.39db**	**28.32%**	**1.39db**
"Bodleian Library" *r*	36.65%	1.98db	46.19%	2.65db	37.61%	2.04db	**53.45%**	**3.21db**
"Radcliffe Camera" *r*	19.95%	0.93db	32.64%	1.64db	25.02%	1.21db	**38.66%**	**2.01db**
"All Souls College" *r*	11.15%	0.50db	23.30%	1.12db	11.80%	0.53db	**38.16%**	**1.88db**



(a) $QP = 22$ (b) $QP = 27$

(c) $QP = 32$ (d) $QP = 37$

Fig. 7: Encoder reference frame decisions by coding unit on "Merton College" for different QP values. Each color corresponds to a reference frame index in the active reference picture set (the projections for our application), while the intra-mode is represented in black.

TABLE III: BD-rate reduction compared with HEVC inter for different methods, with N the number of prediction models.

Sequence	**Zhang [6]**		**Shi [5]**		**Ours**	
	BD-rate	N	BD-rate	N	BD-rate	N
"Merton College"	19.63%	12	24.60%	4	**28.50%**	8
"Bodleian Library"	19.17%	12	27.52%	4	**39.80%**	6
"Radcliffe Camera"	28.82%	12	**42.74%**	4	40.54%	3
"All Souls College"	26.42%	12	44.70%	4	**61.32%**	1
"Holidays-1192"	4.59%	80	1.67%	4	**8.04%**	7
"Holidays-1127"	23.94%	80	33.29%	4	**37.2%**	4
"Merton College" r	13.29%	12	21.17%	4	**28.32%**	8
"Bodleian Library" r	19.80%	12	41.96%	4	**53.45%**	6
"Radcliffe Camera" r	31.02%	12	37.37%	4	**38.66%**	3
"All Souls College" r	31.55%	12	36.50%	4	**38.16%**	1
"Holidays-1192" r	3.8%	80	7.16%	4	**12.59%**	8
"Holidays-1127" r	25.41%	80	27.95%	4	**32.07%**	4
Mean BD-rate gain	20.64%		28.89%		**34.89%**	

B. Comparison to the state of the art

To compare with the state of the art, we implemented the approaches proposed by Shi *et al.* [5] and Zhang *et al.* [6].

The BD-rate gains are reported in Tab. III for the six sequences shown in Fig. 1.

Our method achieves a higher coding performance for the image pairs (a), (b), (d), (e) and (f), with a respective improvement of 3.9%, 12.28%, 16.62%, 3.91% and 3.45%.

Improvements over the state of the art can be explained by the use of a finer prediction model. Restricting the number of models to 4, as proposed by Shi *et al.*, reduces the prediction efficiency as smaller regions could be absorbed into larger ones and thus would not benefit from an accurate prediction model. Zhang *et al.* divide the images into 256x256 pixel "units", which is costly in terms of side information which need to be encoded, this explains the lower performance on all the sequences compared to our scheme. A "unit" can also span over multiple planes and thus results in an incorrect projection estimation. In our scheme, the regions are dependent on the image pair content correlations, thus the models are more robustly estimated from a plain and distinct region, which results in a better prediction. Moreover, the proposed photometric compensation on the color channels can be more efficient than the simple scale-offset compensation model on the luminance channel. For the (c) sequence, our scheme fails to detect the optimal number of models and selects 3 models instead of 4. With a fixed number of 4 models, an improvement of 44.56% over HEVC inter can be obtained. As such, the automatic detection of the number of regions performs well on average, but can result in lower gains on some sequences. Re-learning the parameters of the model fitting on a larger dataset could help improving the performances.

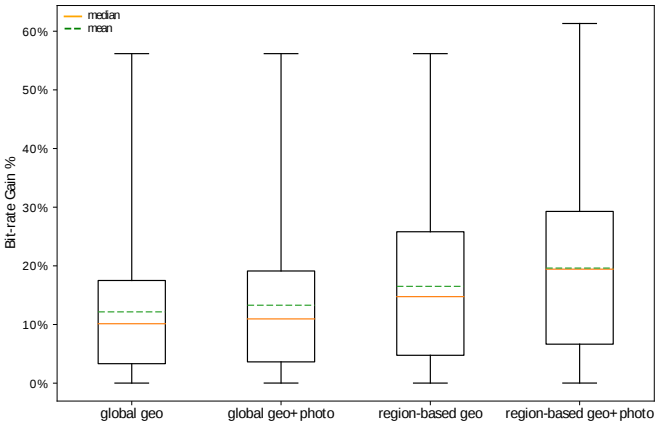


Fig. 8: Overall performance comparison of prediction methods.

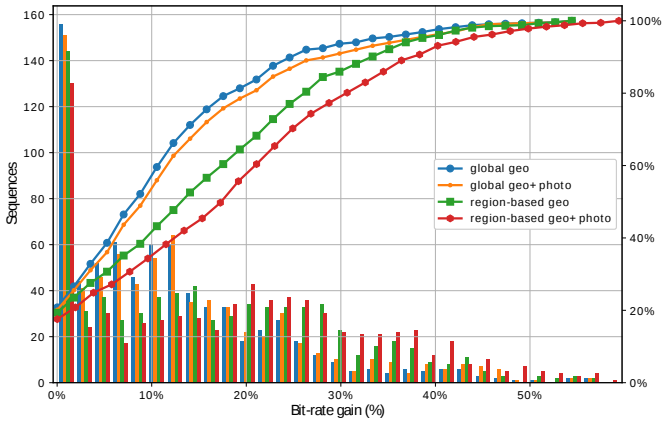


Fig. 9: Distribution of the rate-distortion gains for different prediction methods, with their respective cumulative density function.

C. Overall performance

In this section we present the overall coding performance of the proposed prediction scheme on a large dataset of image

TABLE IV: Mean runtime increases for the total encoding and the HEVC encoding of the proposed scheme, compared to the HEVC inter-coding of two images.

Method	Total	HEVC encoding
global geo	121.54%	110.59%
global geo+photo	120.61%	109.65%
region-based geo	143.17%	125.76%
region-based geo+photo	142.55%	123.45%

pairs. Multiples images were randomly aggregated to form a collection of about 700 sequences from the previously mentioned online databases [29]–[32]. This dataset present a large variety of scene contents, image resolutions and distortions: different cameras, viewpoints, conditions of illuminations, etc… Again, the BD-rate gains reported here are calculated with respect to the performance of the HEVC low-delay inter-coding configuration.

The overall performances in terms of BD-rate saving for different geometric prediction methods, a single model "global" versus our method "region-based", with and without photometric compensation, are shown in Fig. 8.

The mean BD-rate distortion improvements are respectively 12.16%, 13.29%, 16.50% and 19.61%. The distribution of the rate-distortion improvements on the dataset is shown in Fig. 9. The wide range of gain interval (from 0% to 61%) reflects the method high dependency on the inter-image correlation. The full region-based model outperforms the other models, with at least 19.62% gain for half the sequences. While 41.59% of the sequences do not benefit from a photometric compensation, the scale-offset model is selected for 37.61% of the sequences and the piece-wise spline model is more performant on the remaining 20.80%. The very high gains are obtained for frames with a simple global geometric distortion, such as a rotation, which cannot be compensated efficiently by the block motion estimation and compensation of video encoders. The low gains result from either sequences with complex distortions that cannot be compensated with geometric-based compensations (*e.g.* significant optical distortion) or simple distortions already compensated efficiently by block motion compensation. Also our scheme strongly relies on the keypoints extraction and matching step, which can fail for some scenes, as no adaptive method is proposed to control the sensitivity of the detector and the matching threshold. In these cases, no prediction can be performed and thus no improvements over the HEVC "inter" baseline can be expected.

For the BD-rate distortion improvement over HEVC all-intra coding, a mean gain of 21.56% is achieved. The gain of 19.61% obtained over HEVC "inter" indicates that the HEVC inter-prediction models can only handle larger distortions to a limited extent.

D. Complexity study

Compared to a classical pseudo-video coding approach, the main increase in complexity of our scheme resides on the encoder side. The mean encoding run-times of HEVC low-delay, a global prediction scheme and our region-based prediction model have been computed for the same 700 sequences, and are reported in Tab. IV. The mean runtime increase of our scheme is of 142.55% compared with HEVC inter. The increased complexity can be explained both by the overhead of the region-based prediction algorithm and the HEVC inter-coding.

TABLE V: Distribution of the runtime for each step in the region-based scheme.

Step	Runtime ratio
Super-pixels extraction	26.58%
Descriptors extraction and matching	44.86%
Geometric models estimation	2.64%
Geometric models fitting	24.10%
Photometric compensation	1.46%
Misc.	0.36%

TABLE VI: Influence of the "search range" value on the runtime and the rate-distortion performance.

Prediction method	Search Range	Runtime	BD-rate gain
"global"	64px	116.45%	14.44%
"region-based geo"	1px	105.90%	14.98%
	2px	107.18%	15.31%
	4px	107.24%	15.50%
	8px	107.57%	15.69%
	16px	108.76%	15.72%
	32px	111.14%	15.67%
	64px	130.73%	16.50%

The distribution of the increase in complexity of the region-based prediction algorithm is detailed for each step in Tab. V. The slowest step is by far the local descriptors extraction and matching, followed by the super-pixels extraction and the homography models fitting. The SIFT extraction could benefit from a more efficient implementation, such as the GPU one proposed in [35]. The descriptors matching could also be performed on GPU, or leverage the approximate k-nearest-neighbours methods based on KD-Trees such a FLANN [36]. Similarly, an efficient GPU implementation of SLIC have also been proposed [37].

The second significant overhead in the complexity of the proposed scheme is due to the inter-prediction process in HEVC. The 23.45% increase is due to the compensated regions which need to be encoded by HEVC before being available as reference to encode the target frame.

By enforcing multiple reference frames, the encoder has to perform more block motion estimations to compute the potential motion vectors. The default "low-delay" configuration of HEVC sets a search range of 64 pixels for the motion vector, and can be reduced to speed up the encoding at the cost of a reduction of the compression performance. Experimental results are reported in Tab. VI for the 700 sequences. One can observe that by setting a lower value for the motion vector search range, the encoding runtime can be reduced, at the expense of a decreased BD-rate gain, since local geometric disparities would not be well compensated by the constrained block motion compensation. However, it provides a good trade-off between complexity and efficiency.

On the decoder side, the increase in complexity is fairly limited. Once the reference image has been retrieved, only

the additional step of reconstructing the projections from the side information needs to be performed. This step amounts to computing the new pixel coordinates, applying an interpolation and finally correcting the pixel values with the photometric model for each region. This operation has an $O(n)$ complexity, and as such, can be performed in linear time with respect to the image size. With our implementation, it takes less than 1s to generate the reference images on a recent laptop.

VI. Conclusion

In this paper, we presented a novel prediction scheme for cloud-based image compression. Unlike current approaches, our scheme features a semi-local geometric and photometric prediction method able to compensate in a region-wise manner distortions between two images. The proposed scheme can significantly improve the rate-distortion performances compared to classical image and video coding solutions, and is also competitive compared to state of the art methods. The added complexity of our solution is limited and could be reduced by leveraging efficient implementation of the algorithms involved. Furthermore, the proposed prediction method is agnostic to the video codec used, allowing to use existing coding infrastructures without introducing major modifications.

Though we focused in this paper on image sets compression in the cloud, other applications with highly correlated image content such as photo albums compression [38], cloud gaming streaming [39], [40], and traditional video coding [41] could also benefits from the proposed prediction methods.

Interesting challenges still remain, such as exploiting multiple frames from the cloud and ascertaining the scalability of cloud-based image compression techniques. Furthermore, cloud-based image compression solutions rely on classical content-based image retrieval systems, designed for semantic retrieval. Adapting one of these schemes for cloud-based compression applications would provide better references for the prediction and ultimately improve the bit-rate distortion performances.

References

[1] Facebook, Ericsson, and Qualcomm, "A focus on efficiency," *A whitepaper from Facebook, Ericsson and Qualcomm*, September 2013.

[2] G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[3] R. Zou, O. C. Au, G. Zhou, W. Dai, W. Hu, and P. Wan, "Personal photo album compression and management." in *ISCAS*, 2013, pp. 1428–1431.

[4] D. Perra and J. Frahm, "Cloud-scale image compression through content deduplication." in *BMVC*, 2014.

[5] Z. Shi, X. Sun, and F. Wu, "Photo album compression for cloud storage using local features," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 4, no. 1, pp. 17–28, 2014.

[6] Y. Zhang, W. Lin, and J. Cai, "Dense correspondence based prediction for image set compression," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2015, pp. 1240–1244.

[7] 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.

[8] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, May 2001.

[9] M. Douze, H. Jegou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of GIST descriptors for web-scale image search," in *Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR,*.

[10] H. Yue, X. Sun, J. Yang, and F. Wu, "Cloud-based image coding for mobile devices - toward thousands to one compression," *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 845–857, 2013.

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

[12] Z. Yuan, P. Yan, and S. Li, "Super resolution based on scale invariant feature transform," in *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*. IEEE, 2008, pp. 1550–1554.

[13] M. Amintoosi, M. Fathy, and N. Mozayani, "Regional varying image super-resolution," in *Computational Sciences and Optimization, 2009. CSO 2009. International Joint Conference on*, vol. 1. IEEE, 2009, pp. 913–917.

[14] C.-C. Hsu and C.-W. Lin, "Image super-resolution via feature-based affine transform," in *Multimedia Signal Processing (MMSP), 2011 IEEE 13th International Workshop on*. IEEE, 2011, pp. 1–5.

[15] H. Yue, J. Yang, X. Sun, and F. Wu, "Sift-based image super-resolution," in *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 2896–2899.

[16] H. Yue, X. Sun, J. Yang, and F. Wu, "Landmark image super-resolution by retrieving web images," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4865–4878, 2013.

[17] L. Sun and J. Hays, "Super-resolution from internet-scale scene matching," in *Computational Photography (ICCP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1–12.

[18] H. Yue, X. Sun, J. Yang, and F. Wu, "Cid: Combined image denoising in spatial and frequency domains using web images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2933–2940.

[19] ——, "Image denoising by exploring external and internal correlations," *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1967–1982, 2015.

[20] J. Bégaint, D. Thoreau, P. Guillotel, and M. Türkan, "Locally-weighted template-matching based prediction for cloud-based image compression," in *Data Compression Conference, DCC, Snowbird, UT, USA*, 2016, pp. 417–426.

[21] 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.

[22] 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.

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

[24] 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.

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

[26] 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.

[27] H. N. Isack and Y. Boykov, "Energy-based geometric multi-model fitting," *International Journal of Computer Vision*, vol. 97, no. 2, pp. 123–147, 2012.

[28] Y. HaCohen, E. Shechtman, D. B. Goldman, and D. Lischinski, "Optimizing color consistency in photo collections," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 38:1–38:10, 2013.

[29] T. Werner and A. Zisserman, "New techniques for automated architectural reconstruction from photographs," in *Computer Vision - ECCV 2002, 7th European Conference on Computer Vision, Proceedings, Part II*, 2002, pp. 541–555.

[30] H. Jégou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *European Conference on Computer Vision*, ser. LNCS, A. Z. David Forsyth, Philip Torr, Ed., vol. I. Springer, oct 2008, pp. 304–317.

[31] H. Shao, T. Svoboda, and L. V. Gool, "ZuBuD — Zürich buildings database for image based recognition," Computer Vision Laboratory, Swiss Federal Institute of Technology, Tech. Rep. 260, March 2003.

[32] J. Heinly, E. Dunn, and J. Frahm, "Comparative evaluation of binary features," in *Computer Vision - ECCV 2012 - 12th European Conference*

on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II, 2012, pp. 759–773.

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

[34] R. Storn and K. V. Price, "Differential evolution - A simple and efficient heuristic for global optimization over continuous spaces," *J. Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[35] M. Björkman, N. Bergström, and D. Kragic, "Detecting, segmenting and tracking unknown objects using multi-label MRF inference," *Computer Vision and Image Understanding*, vol. 118, pp. 111–127, 2014.

[36] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *International Conference on Computer Vision Theory and Application VISSAPP'09)*. INSTICC Press, 2009, pp. 331–340.

[37] C. Y. Ren, V. A. Prisacariu, and I. D. Reid, "gSLICr: SLIC superpixels at over 250Hz," *ArXiv e-prints*, Sep. 2015.

[38] R. Zou, O. C. Au, G. Zhou, W. Dai, W. Hu, and P. Wan, "Personal photo album compression and management," in *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 1428–1431.

[39] J. Wu, C. Yuen, N.-M. Cheung, J. Chen, and C. W. Chen, "Enabling adaptive high-frame-rate video streaming in mobile cloud gaming applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 12, pp. 1988–2001, 2015.

[40] ——, "Streaming mobile cloud gaming video over tcp with adaptive source–fec coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 1, pp. 32–48, 2017.

[41] 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