Locally-weighted template-matching based prediction for cloud-based image compression

Jean Bégaint*, Dominique Thoreau*, Philippe Guillotel* and Mehmet Türkan†

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

† Izmir University of Economics
Sakarya Cad., No. 156
35330 Izmir, Turkey
firstname.lastname@iue.edu.tr

Abstract

Thanks to the increasing number of images stored in the cloud, external image redundancies can be leveraged to efficiently compress images by exploiting inter-images correlations. In this paper, we propose a novel cloud-based image coding scheme. Unlike current state-of-the-art systems, our method relies on a data dimensionality reduction technique. A global compensation is associated to a locally-weighted template matching compensation method to predict a reference frame, to be then differential-coded with classic video coding tools. Experimental results demonstrate that the proposed approach yields significant rate-distortion performance improvements compared to current image coding solutions.

Introduction

The emergence of cloud applications and web services has led to an increasing use of online resources. A significant part of these services provide storage and distribution solutions for images (*e.g.:* Facebook, Flickr, OneDrive). In 2013, more than 250 billion of pictures had already been uploaded on Facebook and more than 350 were uploaded each day [1]. Due to the amount of data already stored in the cloud, very similar content might already be stored online. By exploiting inter-images redundancies, encoding images together can be more efficient than compressing each image independently with classic image coding such as JPEG [2] or HEVC intra-coding [3]. Several approaches for compressing set of images have thus been studied.

The first works on image sets compression were proposed for medical and satellite imaging. Such images are typically highly correlated as the observation conditions slightly vary. Early solutions involve encoding the target image as its difference to a representative signal (RS) of the images set. The RS can be a reference image, an average or a median of all the images in the set [4]. However, approaches involving the RS are not robust to geometric transformation (*e.g.:* displacements, cropping, focal length) or photometric transformations (*e.g.:* illumination changes). Since the disparities are too steeped, the prediction errors will be too significant to be encoded efficiently, and so, will not bring any bit-rate reduction. Dynamic data-sets are also an issue since the RS has to be recomputed for each modification of the sets, in order to preserve the efficiency of the prediction and the compression.

To exploit the redundancies in a photo album, Zou *et al.* proposed in [5] to leverage the inter-prediction coding tools of HEVC by encoding similar images as pseudo-sequences. The images are organized by correlation in a graph, the similarity

is measured by the Sum of Squared Differences (SSD). Minimum Spanning Trees (MST) are then obtained from the graph and concatenated as pseudo-video sequences, to be then differential-coded using the inter-coding tools of HEVC. A maximum depth of the tree is enforced to control the random access time. This scheme brings significant bit-rate reduction compared to classic coding tools as it makes use of the inter-prediction modes of HEVC to efficiently compress the images. However, the inter-image disparities is measured by the SSD, which is highly sensitive to geometric and illumination changes, the redundancies between the images are thus not fully exploited. Besides, such disparities are not compensated either by video coding tools which mainly deals with 2D-displacements by using Block Motion Compensation (BMC).

Shi *et al.* introduced in [6] a multi-model approach based on the SIFT [7] local feature descriptors. A similar MST approach is used, however the correlation is measured as the distance between groups of matched local descriptors. A feature-based geometric and photometric transformation is then performed to reduce the inter-image redundancies. Finally, the BMC from HEVC is used to compensate for local displacements and reduce local disparities. This approach relies heavily on the clustering of the SIFT features.

Yue *et al.* proposed in [8] an approach to reconstruct an image from local patches fetched from the cloud. The image is reconstructed from patches of correlated content retrieved with the help of the SIFT descriptors and a down-sampled version of the image. Although good visual results and a thousands- to-one compression ratio can be obtained, this technique might not reconstruct faithfully the original image due to the use of sparse local features without local prediction and error coding like traditional image and video coding.

This paper describes a novel cloud-based compression scheme based on a data dimensionality-reduction technique. Assuming a correlated image can be found in the cloud, a global geometric and photometric compensation is first performed to account for the geometric transformations and illumination disparities. A locally-weighted template-matching based prediction is then performed to predict a reference image. Finally, the target image is inter-encoded with the predicted reference image in HEVC. A crucial aspect of this scheme is the accurate prediction that can not be achieved by coarser global compensation mechanisms based on sparse local features. Thus, we can leverage the advanced inter-coding modes of video encoders such as HEVC.

The rest of this paper is organized as follows. Section 2 gives an overview of the proposed compression scheme. Section 3 explains it in details. Experiments are reported Section 4. Section 5 concludes this paper.

Overview of the proposed scheme

The framework of our compression scheme can be decomposed in two steps, as shown in Fig. 1. For the purpose of explanation, we will only consider a pair of similar images but the scheme can also be adapted for larger sets of similar images. When considering an image (I_c) to be encoded, a reference image (I_r) is first retrieved from

Figure 1: Proposed compression framework

the cloud with the help of a classic Content Based Image Retrieval (CBIR) system. A predicted image (I_p) is then constructed from the inter-images correlations. The current image I_c is then coded in reference to the predicted image I_p with HEVC. The main purpose here is to predict I_c with sufficient accuracy so as to leverage the efficient inter-prediction modes of HEVC and thus reduce the bit-rate required to encode I_c. In this way, a new image uploaded online can be encoded more efficiently from a similar image retrieved from the cloud than with static picture coding tools.

The prediction mechanism proposed in this paper relies on a global compensation which is associated to a local prediction mechanism, to compensate accurately the multiple disparities that may occurred between the images. The global compensation is performed to adjust the geometric and photometric distortions on the global level. A template-matching based method is then applied on the pixels at block level to compensate local distortions.

The predicted frame is used as a reference frame for the compression with HEVC. The current image I_c and the predicted frame I_p are concatenated in a pseudo-video sequence and encoded as a classic video (with both intra and inter coding modes enabled). The side information required to reconstruct the reference frame on the decoder side is also transmitted and taken into account in the bit-rate. For the decoding, the predicted image is reconstructed from the reference image and then used to decode the current image. 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.

Proposed cloud-based compression scheme

Global compensation mechanism

The geometric deformation between two similar images can be described as an homography, which models the perspective transformation between two planes (Eq. 1), where $h_{11}, h_{22}, \ldots, h_{32}$ are homography coefficients.

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \tag{1}$$

To compute the transformation model one widely used approach consists in estimating the transformation from matched local descriptors. Local descriptors are first extracted on both images and then matched in pairs. We used the SURF descriptors [9] to describe the local features in our implementation. A transformation model is then estimated from the matched descriptors by applying the RANSAC [10] algorithm. The process is shown Fig. 2.

To compensate for the photometric disparities, we used the global color mapping algorithm proposed in [11]. A correction method [12] for panorama and an histogram reshaping technique [13] were first considered but the adjustments suffer from some inconsistencies if some colors are not present in both images and also can not handle saturation. The method consists in fitting each RGB channel on a constrained monotonic curve and then apply a linear transform to prevent saturation. The saturation adjustment is modeled with a multiplying coefficient s on the gray line. The matrix C (Eq. 2) is used to solve the saturation problem with several weighting models for w_r, w_g, w_b (uniform and YUV) and the best result is kept. This method present several advantages as it is robust to saturation, non-linear tone adjustments and can handle colors that do not overlap in the image pairs [11].

$$C = \begin{bmatrix} s - w_r & w_g & w_b \\ w_r & s - w_g & w_b \\ w_r & w_g & s - w_b \end{bmatrix} \tag{2}$$

The homography matrix and the photometric adjustment parameters are then encoded and included in the bit-stream for further decoding, only the 2 matrices H and C need to be encoded, regardless of the size of the image.

Locally-weighted template-matching based prediction

In this section, we describe an inter prediction algorithm based on Locally Linear Embedding (LLE) [14]. Specifically we used the locally-weighted template-matching prediction algorithm introduced in [15] as an inter prediction mode for video compression such as in H.264/HEVC.

The Template Matching (TM) method is derived from the block matching technique used in BMC. Instead of looking for the most correlated blocks in the reference frame, TM exploits the correlation between a block and its neighboring pixels, the template (Fig. 3). As opposed to block matching, there is usually no need to transmit a motion vector since the template of the current block is known on the decoder side (the templates being composed of the previously decoded neighboring blocks).

LLE, a data dimensionality reduction technique, has been proposed for intra-prediction in [16, 17] as an extension to TM. The main idea consists in representing a data point of a high-dimensional space as an linear approximation of its neighbors.

The inter prediction method presented here relies on the assumption that each block in the current frame can be approximated from correlated blocks retrieved in a similar frame. Each block in the current image I_c is thus reconstructed as a linear combination of its K Nearest Neighbors (K-NN) retrieved from the reference image I_f.

(a) Current image (b) Reference image (c) Keypoints Matching

(d) Inliers selection (e) Compensated image

Figure 2: Global Compensation

First, the nearest-neighbor block is retrieved from the reference frame. Then the $K-1$ nearest neighbors of the previously retrieved block are fetched in the reference frame (cf. Fig. 3). The matching is performed on the patch level, *i.e.:* the block and its template. By using the texture information of the current block, a better prediction is obtained, as a block and its template might not be sufficiently correlated.

Only the luminance is considered for measuring the euclidean distance between the patches. As we are using the default 4:2:0 color sampling of HEVC, the luminance is the most significant component for the search. By performing the K-NN search in two steps, only the position of the first NN block in the reference frame has to be transmitted on the decoder side. Since the position of the reference block is relative to the position of the current block, encoding its position comes down to encoding a motion vector, a task at which classic video coding tools are very efficient.

The current template is then estimated from the retrieved templates by solving a constrained least square approximation (Eq. 3). The vector X, composed of the pixels of the current template, is approximated as a linear combination of A, composed of the pixels of the K-NN templates, and V, the LLE weights. V is obtained by solving first $DV = 1$, with D the local covariance matrix with respect to X, then the weights are normalized to sum to one.

$$\min_{V} ||X - AV||_2^2 \quad \textbf{s.t.} \quad \sum_{k=1}^{K} V_k = 1 \tag{3}$$

Figure 3: K-NN search in the reference image using the template-matching

Figure 4: Examples of images contained in our dataset.

From the retrieved K nearest neighbors patches, the LLE coefficients are computed from the templates, for each channel (Y,U,V). The current block can then be reconstructed from the retrieved blocks. By computing the coefficients from the templates instead of the blocks, there is no need to store the coefficients in the bit-stream as they are already available on the decoder side from the previously decoded blocks, which strongly increases the efficiency of the encoding. The additional cost of this local prediction amounts therefore for only one motion vector per block.

Experiments

The experiments were conducted on multiple pairs of similar images presenting challenging disparities aggregated from publicly available databases [18–20], some examples are shown in Fig. 4. These images display a combination of differences of viewpoint, focal length, illumination, rotations. The HEVC HM 16.2 software was used for the video encoding. For the JPEG encoding we use the ImageMagick library (6.9.0-Q16). The rate-distortion performances presented in this paper are computed

(a) RD performance w.r.t. the precision ($k = 128$)

(b) RD performance w.r.t. K (*Qpel* precision)

Figure 5: Rate-distortion (RD) performance comparison

	$p = 1$		$p = 0.5$		$p = 0.25$	
K	PSNR (dB)	Bitrate (%)	PSNR (dB)	Bitrate (%)	PSNR (dB)	Bitrate (%)
2	-0.25	2.90	0.28	-4.42	2.66	-37.61
4	0.03	-0.35	0.57	-7.91	3.45	-47.66
8	-0.08	0.71	0.58	-7.86	3.73	-49.74
16	-0.01	0.16	0.76	-10.14	5.50	-64.01
32	-0.01	0.13	1.11	-14.76	6.93	-71.66
64	0.04	-0.46	1.56	-19.90	7.50	-74.03
128	0.13	-1.52	1.76	-22.14	7.78	-75.25

Table 1: Average bit-rate and PSNR gains compared to HEVC inter-coding

with the Bjonntegard metrics [21], using the following settings for the Quantization Parameter (QP): 22, 27, 32 and 37. For the experiments, we used a block size of 8×8 pixels and 3 pixels wide template, composed of the upper and left pixels. Only the PSNR of the luminance channel is reported here.

The principal drawback of the LLE method is the high complexity of the algorithms involved. The main bottleneck is the K-NN search which scales in $O(DN^2)$, with D the dimension of the vector and N the number of data points. Several suggestions were proposed in [14] to reduce the complexity of the method. The complexity can be reduced by using a K-D Tree structure, a binary tree designed for efficiently organizing data points in a k-dimensional space. K-D Trees can thus be used for high-dimensional nearest-neighbors searches. The complexity of the K-NN search drops to $O(DNlog(N))$. Several implementations are publicly available, for our experiments we used the FLANN library [22].

Although images are sampled at the pixel level, the accuracy of block-based geometric transformations can be improved by working at sub-pixels levels. The template-matching based algorithm was thus tested with different precisions p: pixel,

(a) Original Image (b) *pel* precision (21.18 dB PSNR) (c) *half-pel* precision (24.65 dB PSNR) (d) *Qpel* precision (29.92 dB PSNR)

Figure 6: Visual quality comparison ($K = 64$).

half-pel, and quarter-pel (*Qpel*). Also the complexity for the K-NN search being quadratic, or linearithmic at best with approximated search methods, determining the optimal number of neighbors to retrieve is critical. Several values of K ($K \in \{2^i \mid i \in [1, 7]\}$) were tested at different precisions. The results obtained for 30 pairs of images, different sub-pel accuracies and K values are presented Table 1. The best performances are undeniably obtained at *Qpel* precision, where using a K value greater than 64 does not bring significant improvements (Fig. 5b, Fig. 6). Also it may be observed that the order of magnitude of the gain is restricted by the (sub)-pixel precision. Different codecs and versions of our scheme are compared in Fig. 5a.

By exploiting the inter-modes of HEVC with the predicted image, our scheme can improve the coding efficiency up to 75% bit-rate saving compared to classic HEVC inter-coding, and 97% compared to intra image coding (JPEG) (Fig. 5a). While the complexity of the prediction is important, the pixel precision and the number of neighbors K can be set to constrain the processing time at the expense of the encoding performance.

Conclusion

In this paper, we present a novel scheme for cloud-based image compression. Unlike previous approaches, our scheme features a global compensation associated to a local weighted template-matching as a prediction method. By leveraging correlated images already stored in the cloud, our scheme is able to encode efficiently newly uploaded images and brings significant improvements compared to classic image coding tools.

The prediction currently resides outside the HEVC encoder, an implementation in the reference HEVC HM software could be considered to take advantage of the Quadtree-based partitioning and the advanced Rate Distortion Optimization (RDO). From another perspective, the current scheme allows to use the existing coding infrastructures without introducing major modifications. The proposed scheme has also some limitations regarding the high complexity of the algorithms involved. Reducing the complexity of the coding scheme constitutes an important future work.

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] 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.

[4] K. Karadimitriou, "Set redundancy, the enhanced compression model, and methods for compressing sets of similar images," Ph.D. dissertation, Department of Computer Science, Louisiana State University, Baton Rouge, La, USA, 1996.

[5] 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.

[6] 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.

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

[8] 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.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[10] 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.

[11] Y. HaCohen, E. Shechtman, D. B. Goldman, and D. Lischinski, "Non-rigid dense correspondence with applications for image enhancement," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 70:1–70:10, Jul. 2011.

[12] Y. Xiong and K. Pulli, "Color correction for mobile panorama imaging," in *The First International Conference on Internet Multimedia Computing and Service, ICIMCS '09, Kunming, Yunnan, YT, China, November 23-25, 2009*, 2009, pp. 219–226.

[13] T. Pouli and E. Reinhard, "Progressive color transfer for images of arbitrary dynamic range," *Computers & Graphics*, vol. 35, no. 1, pp. 67–80, 2011.

[14] L. K. Saul and S. T. Roweis, "An introduction to locally linear embedding," Tech. Rep., 2000.

[15] M. Alain, S. Cherigui, C. Guillemot, D. Thoreau, and P. Guillotel, "Locally linear embedding methods for inter image coding," in *IEEE International Conference on Image Processing, ICIP 2013, Melbourne, Australia, September 15-18, 2013*, 2013, pp. 1904–1908.

[16] M. Türkan and C. Guillemot, "Image prediction based on neighbor-embedding methods," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1885–1898, 2012.

[17] S. Cherigui, C. Guillemot, D. Thoreau, P. Guillotel, and P. Pérez, "Map-aided locally linear embedding methods for image prediction," in *19th IEEE International Conference on Image Processing, ICIP 2012, Lake Buena Vista, Orlando, FL, USA, September 30 - October 3, 2012*, 2012, pp. 2909–2912.

[18] 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. [Online]. Available: http://lear.inrialpes.fr/pubs/2008/JDS08

[19] 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.

[20] 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. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33709-3_54

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

[22] 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.