QUANTIZATION BIN MATCHING FOR CLOUD STORAGE OF JPEG IMAGES

Xianming Liu[1,2], Gene Cheung[2], Chia-Wen Lin[3], Debin Zhao[1]

[1]School of Computer Science and Technology, Harbin Institute of Technology, China
[2]National Institute of Informatics, Japan
[3]Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan

ABSTRACT

Social media sites like Facebook are obligated to store all photos uploaded by an ever growing user base—which translates to an increasingly expensive storage cost—but only a fraction of uploaded images are revisited thereafter. In this paper, we propose a cloud storage system that trades off computation of a small fraction of requested images with storage of all photos. The key idea is to re-encode uploaded JPEG photos with coarser quantization parameters (QP) for permanent storage, then exploit a signal sparsity prior during inverse mapping to recover fine quantization bin indices via a maximum a posteriori (MAP) formulation. Because by design the system guarantees recovery of an original compressed image (either with exactly the same input fine quantization bin indices or has visual quality indistinguishable by human eyes), from the user's viewpoint it is a normal cloud storage, while from the operator's viewpoint there is pure compression gain and hence lower storage cost. Experimental results show that our storage system can reap significant storage savings (up to 20%) at roughly the same image PSNR (within 0.13dB).

Index Terms— cloud storage, image compression, sparse signal recovery

1. INTRODUCTION

The popularity of social media sites like Facebook and photo sharing sites like Flickr means that as the user base continues to grow, the responsible operators are on the hook to store an increasingly large number of photos uploaded by users[1]. However, because human's time and attention are fundamentally limited resources despite the ever growing heap of information generated [1], only a fraction of the uploaded images would realistically be revisited thereafter. This poses a system implementation challenge for photo hosting operators: how to best exploit this asymmetry between the volumes of initial uploaded images and actual requested photos to minimize the overall storage cost?

In this paper, we propose an efficient cloud storage system that trades off computation of a small number of requested images with storage of all photos. The key idea is the following: when a *user* uploads a JPEG image compressed using quantization parameters (QP) $\mathbf{q}$, the serving *cloudlet* first re-encodes the image using coarser QP $\mathbf{Q}$. The re-encoded image is stored at a *central cloud* for permanent storage. If / when the image is requested, the coarsely quantized image is retrieved by the cloudlet, who then performs a reverse mapping to recover fine quantization bin indices with the help of a signal prior. The restored image is returned to the user. The system is illustrated in Fig. 1. From the user's viewpoint, the retrieved image is the same compressed image as the one uploaded (either with exactly the same input fine quantization bin indices or has visual quality indistinguishable by human eyes), hence it is no different than typical cloud storage. From the operator's viewpoint, however, the more coarsely quantized image due to re-encoding results in pure compression gain and lower storage cost.

The crux of the system rests in the inverse mapping from coarse quantization bin indices to fine quantization bin indices at the cloudlet when an image is requested: we call this the *quantization bin matching* (QBM) problem. For each fixed-size N-pixel code block, we formulate a *maximum a posteriori* (MAP) problem to find the most probable N fine quantization bin indices in the frequency domain given N coarse quantization bin indices, assuming a sparse signal prior [2], where a signal well approximated by a sparse linear combination of dictionary atoms is more likely. This formulation requires integration of the prior distribution over a given N-dimensional quantization cell, which is mathematically more challenging than previous de-quantization work [3, 4] where a single most probable signal within a quantization cell is sought. We propose a fast algorithm to approximate the prior distribution integration. Experimental results show that our storage system can reap significant storage savings (up to 20%) at roughly the same image PSNR (within 0.13dB).

The outline of the paper is as follows. We first discuss related work in Section 2. We overview our cloud storage system in Section 3. We formulate our MAP quantization bin matching problem in Section 4 and present our algorithm in Section 5. Finally, experimental results and conclusion are presented in Section 6 and 7, respectively.

[1] It is estimated in 2013 that 350 million photos are uploaded to Facebook every day.

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

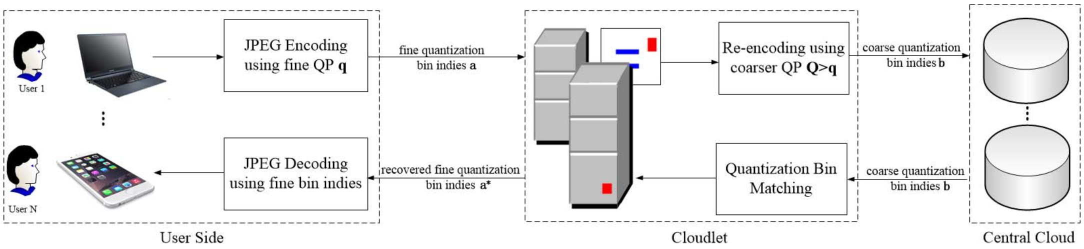
User 1
User N
JPEG Encoding using fine QP **q**
JPEG Decoding using fine bin indies
fine quantization bin indies **a**
recovered fine quantization bin indies **a***
Re-encoding using coarser QP **Q>q**
Quantization Bin Matching
coarse quantization bin indies **b**
coarse quantization bin indies **b**
User Side
Cloudlet
Central Cloud

Fig. 1. Block diagram of the proposed scheme.

2. RELATED WORK

Image compression for cloud storage has attracted much interest in both industry and academia. For example, Yue *et al.* proposed in [5] to describe images and reconstruct them from a large-scale image database via the SIFT descriptor. Shi *et al.* [6] further extended this idea to compress a photo album by exploiting local features rather than pixel values for analyzing and exploring the correlation among images. Song *et al.* [7] proposed a novel cloud-based distributed image coding scheme. In this method, an input image is reconstructed in the cloud using retrieved correlated images, from which side information (SI) is extracted. The image is then compressed through a transform-domain syndrome coding to correct the disparity between the original image and the SI by an iterative refinement process. In [8], Perra *et al.* proposed to take advantage of redundant image data in the cloud by independently compressing each newly uploaded image with its GIST nearest neighbor taken from a canonical set of uncompressed images. Although these schemes achieve good coding gain, they rely heavily on whether correlated images can be found in the cloud. Moreover, these methods cannot guarantee exact recovery of the original photos uploaded by users.

3. SYSTEM OVERVIEW

For simplicity, we assume the following simple interaction among the three entities—a *user*, a *cloudlet* and a *central cloud*—in the cloud storage system illustrated in Fig. 1. A user uploads a compressed image, coded using JPEG that computes for each N-pixel code block m a set of DCT quantization indices $\mathbf{a}(m) = [a_1(m), \ldots, a_N(m)]$ using quantization parameters (QP) $\mathbf{q} = [q_1, \ldots, q_N]$, to a serving cloudlet. The cloudlet in turn *re-encodes*[2] the image using coarser QP $\mathbf{Q} = [Q_1, \ldots, Q_N]$ that maps each $\mathbf{a}(m)$ to corresponding quantization index set $\mathbf{b}(m)$ for code block m, and stores them in the central cloud. When the user requests the image, the cloudlet retrieves the index sets $\mathbf{b}(m)$'s, performs an inverse mapping that converts each $\mathbf{b}(m)$ back to $\mathbf{a}(m)$, and returns $\mathbf{a}(m)$'s back to the user. Because QP $\mathbf{Q}$ are coarser than original QP $\mathbf{q}$, *i.e.*, $Q_i \geq q_i$, there tend to be more zeros in $b_i(m)$'s than $a_i(m)$'s, resulting in compression gain when suitable entropy coding is employed. Our system is designed to ensure that the input image can be satisfactorily recovered, which means that either: i) original $\mathbf{a}(m)$ for each block m is recovered exactly, or ii) the recovered image is so close to the input compressed image in PSNR that to the user they are visually indistinguishable.

The technical challenge is to design an inverse mapping $f(\mathbf{b})$, so that for some QP $\mathbf{Q}$ coarser than original $\mathbf{q}$ (thus resulting in coding gain), the input image can be recovered well. We only require that this inverse mapping is possible for a large enough fraction of blocks in the image that are re-encoded with QP $\mathbf{Q}$; the remaining blocks are unchanged with QP $\mathbf{q}$. A small binary image losslessly encoded with JBIG [9] is encoded in addition to inform the decoder which blocks are re-quantized using $\mathbf{Q}$. We discuss this coarse-to-fine bin matching problem in the following section.

4. QUANTIZATION BIN MATCHING

4.1. MAP Formulation for Quantization Cell

Denote by X_i the ground-truth i-th DCT coefficient of a target code block $\mathbf{x}$. Correspondingly, denote by a_i the assigned quantization bin index given X_i from compression at the user end, using QP q_i, *i.e.*,

$$a_i = \text{round}\left(X_i/q_i\right). \tag{1}$$

Thus at the decoder with a_i and q_i we know that X_i must reside in interval $I(a_i, q_i) = [(a_i - 0.5)q_i, (a_i + 0.5)q_i)$.

Denote by b_i the assigned quantization bin index for X_i in the re-encoded version using QP Q_i, where $Q_i \geq q_i$. After the re-encoding, X_i must also reside in a larger interval $I(b_i, Q_i) = [(b_i - 0.5)Q_i, (b_i + 0.5)Q_i)]$. Because $I(a_i, q_i)$ in the original encoding and $I(b_i, Q_i)$ in the re-encoding both contain X_i, they must overlap. Thus when performing the inverse mapping $f : b_i \longmapsto a_i$, we only need to consider the *feasible bin set* $\mathcal{F}_{b_i}$:

$$\mathcal{F}_{b_i} = \{a_i \mid I(a_i, q_i) \cap I(b_i, Q_i) \neq \emptyset\} \tag{2}$$

Denote by $P(X_i)$ the prior probability of X_i. We follow a *maximum a posteriori* (MAP) formulation, where we seek the most likely bin $\hat{a}_i$ among a discrete set $\mathcal{F}_{b_i}$:

$$\hat{a}_i = \arg\max_{a_i \in F_{b_i}} \int_{(a_i - 0.5)q_i}^{(a_i + 0.5)q_i} P(X_i \,|\, b_i)\, dX_i \tag{3}$$

[2] The mapping from $\mathbf{a}(m)$ to $\mathbf{b}(m)$ may not be unique; one can perform a de-quantization procedure [3] to estimate the original signal, and then identify coarse quantization bin indices $\mathbf{b}(m)$ that contain the estimated signal.

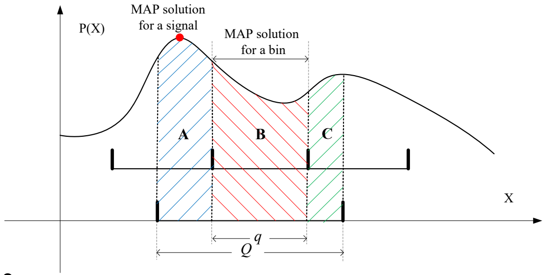

Fig. 2. Illustration of the QBM problem (1D case), where three fine quantization bins (of width q) overlap with the one coarse quantization bin (of width Q).

where

$$P(X_i \,|\, b_i) = \begin{cases} P(X_i) & \text{if } X_i \in I(b_i, Q_i) \\ 0 & \text{o.w.} \end{cases} \tag{4}$$

This MAP formulation of the QBM problem in 1D is illustrated in Fig. 2, where we choose, among the three fine quantization bins of width q that overlap with the coarse quantization bin of width Q, the one with the largest *aggregate* probability $\int P(X_i|b_i)\, dX_i$. This differs from typical de-quantization work like [3], where the MAP formulation for the most probable signal within a quantization bin will simply lead to the peak location of $P(X_i|b_i)$ within the bin.

Considering now all N DCT coefficients $\mathbf{X}$ in a block $\mathbf{x}$ together with N-dimensional prior probability $P(\mathbf{X})$, we formulate the following more general MAP optimization problem:

$$\max_{\mathbf{a}|a_i \in F_{b_i}} \int_{(a_1-0.5)q_1}^{(a_1+0.5)q_1} \cdots \int_{(a_N-0.5)q_N}^{(a_N+0.5)q_N} P(\mathbf{X} \,|\, \mathbf{b})\, dX_1 \ldots dX_N \tag{5}$$

where

$$P(\mathbf{X} \,|\, \mathbf{b}) = \begin{cases} P(\mathbf{X}) & \text{if } X_i \in I(b_i, Q_i), \;\; 1 \leq i \leq N \\ 0 & \text{o.w.} \end{cases} \tag{6}$$

We can rewrite (5), which involves multi-dimensional integration, into a simpler form:

$$\max_{\mathbf{a}|a_i \in F_{b_i}} \bar{P}_{\mathbf{a}}(\mathbf{X}|\mathbf{b}) \tag{7}$$

where $\bar{P}_{\mathbf{a}}(\mathbf{X}|\mathbf{b})$ is the aggregate of probability $P(\mathbf{X}|\mathbf{b})$ within a quantization cell $\mathcal{C}_{\mathbf{a}}$ defined by $\mathbf{a}$, *i.e.*, $I(a_i, q_i), i = 1, \ldots, N$. The challenge to solve (7) is thus how to appropriately define $P(\mathbf{X})$, and how to compute $\bar{P}_{\mathbf{a}}(\mathbf{X}|\mathbf{b})$ efficiently and accurately.

4.2. Defining Prior Probability $P(\mathbf{X})$

We assume a sparsity model [2] to define signal prior $P(\mathbf{X})$. Specifically, a k-sparse signal model states that an N-dimensional signal $\mathbf{x}$ in the pixel domain can be well approximated by a linear combination of k or fewer atoms from an over-complete dictionary $\mathbf{\Phi}$:

$$\mathbf{x} = \mathbf{\Phi}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \quad \|\boldsymbol{\alpha}\|_0 \leq k, \tag{8}$$

where the model error $\boldsymbol{\varepsilon}$ is small. In this paper, following the work [10], we learn adaptive dictionaries via PCA in a machine-learning driven manner. By applying DCT operator $\mathbf{T}$ on both sides of (8), we can get

$$\mathbf{X} = \mathbf{\Psi}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}', \tag{9}$$

where $\mathbf{\Psi} = \mathbf{T}\mathbf{\Phi}$. Therefore, the sparsity of $\mathbf{X}$ can be measured by the sparsity of $\mathbf{x}$ with respect to $\mathbf{\Phi}$.

Given this model, the joint probability distribution $P(\mathbf{X})$ can be represented as

$$P(\mathbf{X}) = \exp\left\{-\|\boldsymbol{\alpha}\|_0/\sigma\right\}, \tag{10}$$

where σ is a model parameter. We further relax the ℓ_0 norm to ℓ_1 norm. Finally, the objective function can be rewritten as

$$\begin{aligned} &\arg\min_{\mathbf{a}|a_i \in F_{b_i}} \textstyle\int_{(a_1-0.5)q_1}^{(a_1+0.5)q_1} \cdots \int_{(a_N-0.5)q_N}^{(a_N+0.5)q_N} \|\boldsymbol{\alpha}\|_1 \, dX_1 \ldots dX_N, \\ &\text{s.t.,} \quad \mathbf{x} = \mathbf{\Phi}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}. \end{aligned} \tag{11}$$

As stated above, the QBM problem can be reformulated as finding the quantization cell with the most and sparsest solutions within its boundaries, rather than a single best sparse solution which is usually done in image de-quantization [3, 4].

5. OPTIMIZATION

Optimizing the objective (11) directly is difficult. Instead of searching for all possible sparse solutions within a quantization cell, we find a single best sparse solution as a representative, then multiply the prior probability of the solution (exponential of the solution's sparsity count) by the cell's volume. This is an approximation of the multi-dimensional integration in (11); in essence we say that the larger the cell volume, the more likely we will find other sparse solutions like the discovered one, hence a larger aggregate probability.

The problem of finding an initial sparse solution within the indexed coarse quantization cell $\mathbf{b}$ is:

$$\begin{aligned} &\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{x} - \mathbf{\Phi}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \\ &\text{s.t.,} \quad X_i \in I(b_i, Q_i), i = 1, \cdots, N, \text{and } \mathbf{X} = \mathbf{T}\mathbf{x}. \end{aligned} \tag{12}$$

The optimization for sparse solution $\boldsymbol{\alpha}^*$ can be effectively and efficiently solved by a fast ℓ_1-minimization algorithm, known as *Augmented Largrangian Methods* (ALM) [11].

We identify the fine quantization cell $\mathbf{a}^*$ that contains this sparse solution as follows:

$$\mathbf{a}^* = \text{round}\left(\mathbf{T}\mathbf{\Phi}\boldsymbol{\alpha}^*/\mathbf{q}\right). \tag{13}$$

If the identified fine quantization cell $\mathbf{a}^*$ overlaps only partially with the coarse quantization cell $\mathbf{b}$ (such as cell A

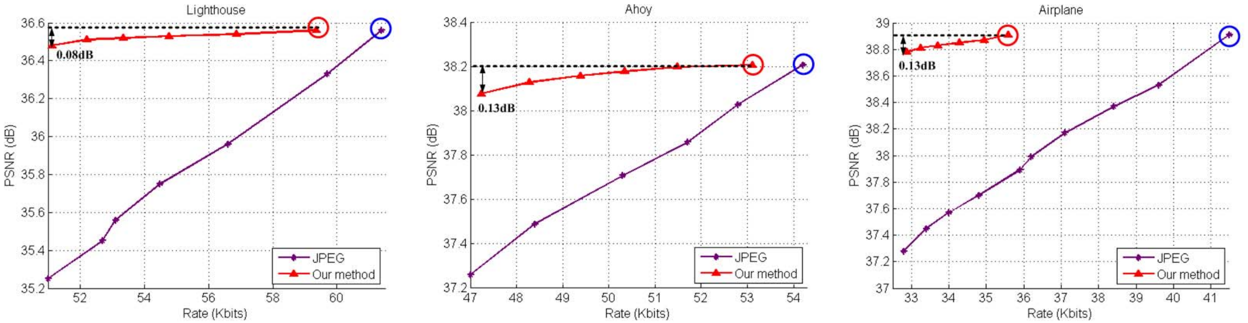

Fig. 3. Rate-distortion performance comparison. The blue circles are the RD points of JPEG compression using fine QF. The red circles represent the RD points of signal fidelity reconstruction by our method. The red and blue circles have the same PSNR values, but red ones use less bits.

Table 1. Pure compression gain without changing uploaded images

Images	Fine QF	Coarse QF	Blocks Proportion	Bit-Saving
Lighthouse	80	50	10.97%	3.32%
Sailboats	80	50	10.09%	2.08%
Window	80	50	12.81%	3.04%
Girl	80	50	14.71%	4.54%
Hats	80	50	15.67%	4.38%
Parrots	80	55	17.37%	3.68%
Airplane	80	55	42.18%	14.19%
Ahoy	80	50	7.96%	2.01%

and C in Fig. 2), its aggregate probability is small since the cell volume is small. We then search for sparse solutions also in adjacent fine quantization cells via (12) to test other solution candidates. Among these cells, the one with the largest product of signal prior (of the identified sparse solution in the cell) and cell volume is concluded as the final solution.

6. EXPERIMENTS

We conducted experiments to demonstrate the effectiveness of our proposed cloud image storage scheme. We use the well-known Kodak dataset, in which five images are randomly selected as training data for dictionary learning, and the rest are used as test images. The quality factor (QF) in JPEG coding, ranging from 1 to 100, indicates the relative visual quality of an image after compression by choosing one from a set of quantization matrices. In our experiments, the fine QF for JPEG coding is fixed at 80, while the coarse QF is chosen from 50 and 55 for different images.

The first design objective is to guarantee exact recovery of fine quantization bin indices as the original uploaded version. In this case, exact bin index recovery means there is no loss in quality, and hence we report only bit savings. Note that bit saving is achieved by re-encoding a fraction of image blocks with coarser QFs that can guarantee exact bin index recovery. Table 1 tabulates the used fine and coarse QFs, the proportion of blocks selected to be re-encoded, and the pure compression gains of our method on the eight test images. The results show that, for the test images, up to 42.18% blocks are selected for further compression using coarse QFs, translating to up to 14.19% bit-saving.

The second alternative design objective is to guarantee recovery of an uploaded image so close in quality that it is indistinguishable to human eyes. To accomplish this, we relax the "exact bin matching" constraint by allowing some differences between the recovered and the input fine bin indices. The level of difference is measured by the sum of block-level bin errors. In our experiments, we test six cases: the sum of error is 0, 2, 3, 4, 5 and 6. Fig. 3 illustrates the rate-distortion performance comparison of our method with JPEG on three test images: *Lighthouse*, *Ahoy* and *Airplane*. The PSNR losses in the six cases are within 0.13dB. It is widely accepted that such small PSNR loss typically does not lead to visual differences. The results show that our method significantly outperforms JPEG. Further, at the expense of very slight PSNR loss, the visually indistinguishable reconstruction option achieves significantly higher bit-saving, compared with the case of exact bin index recovery. Using the error sum of 6 as an example, our method achieves 16.71%, 12.82% and 20.7% bit-saving for *Lighthouse*, *Ahoy* and *Airplane*, respectively, in contrast to 3.32%, 2.01% and 14.19% shown in Table 1.

7. CONCLUSION

We propose a cloud storage system for JPEG images that trades off computation of a small number of requested images with storage of a much larger volume of uploaded images. Specifically, given an input JPEG image uploaded by a user and quantized using quantization parameter (QP) $\mathbf{q}$, a cloudlet re-encodes it using a coarser QP $\mathbf{Q}$ for permanent storage. When the image is requested, the cloudlet retrieves the coarsely quantized image and performs a coarse-to-fine bin matching to recover the input fine quantization bin indices. Experiments show that at virtually the same input image quality, our system can reap significant storage saving.

8. ACKNOWLEDGEMENTS

This work is supported by the Major State Basic Research Development Program of China (973 Program 2015CB351804) and the National Science Foundation of China under Grants 61300110.

9. REFERENCES

[1] D. J. Levitin, *The Organized Mind: Thinking Straight in the Age of Information Overload*, Penguin, 2014.

[2] M. Elad, M.A.T. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proc. IEEE*, vol. 98, no. 6, pp. 972–982, June 2010.

[3] X. Liu, G. Cheung, X. Wu, and D. Zhao, "Inter-block soft decoding of JPEG images with sparsity and graph-signal smoothness priors," in *IEEE International Conference on Image Processing*, Quebec City, Canada, September 2015.

[4] X. Liu, X. Wu, J. Zhou, and D. Zhao, "Data-driven sparsity-based restoration of JPEG-compressed images in dual transform-pixel domain," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

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

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

[7] X. Song, X. Peng, J. Xu, G. Shi, and F. Wu, "Cloud-based distributed image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 6, pp. 1–1, June 2016.

[8] D. Perra and J. Frahm, "Cloud-scale image compression through content deduplication," in *Proc. British Machine Vis. Conf.* 2014, BMVA Press.

[9] F. Ono, W. Rucklidge, R. Arps, and C. Constantinescu, "JBIG2–the ultimate bi-level image coding standard," in *IEEE International Conference on Image Processing*, Vancouver, Canada, September 2000.

[10] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, July 2011.

[11] A.Y. Yang, Zihan Zhou, A.G. Balasubramanian, S.S. Sastry, and Yi Ma, "Fast l_1-minimization algorithms for robust face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3234–3246, Aug 2013.