ANALYZING SENSOR QUANTIZATION OF RAW IMAGES FOR VISUAL SLAM

Olivia Christie[⋆] *Joshua Rego*[⋆] *Suren Jayasuriya*[⋆†]

[⋆] School of Electrical, Computer and Energy Engineering, Arizona State University
[†] School of Arts, Media and Engineering, Arizona State University

ABSTRACT

Visual simultaneous localization and mapping (SLAM) is an emerging technology that enables low-power devices with a single camera to perform robotic navigation. However, most visual SLAM algorithms are tuned for images produced through the image sensor processing (ISP) pipeline optimized for highly aesthetic photography. In this paper, we investigate the feasibility of varying sensor quantization on RAW images directly from the sensor to save energy for visual SLAM. In particular, we compare linear and logarithmic image quantization and show visual SLAM is robust to the latter. Further, we introduce a new gradient-based image quantization scheme that outperforms logarithmic quantization's energy savings while preserving accuracy for feature-based visual SLAM algorithms. This work opens a new direction in energy-efficient image sensing for SLAM in the future.

Index Terms— Visual SLAM, image sensor quantization, RAW images, embedded computer vision

1. INTRODUCTION

Simultaneous localization and mapping (SLAM) is one of the most critical algorithms for robotic and embedded platforms performing navigation in the real world. SLAM, using a combination of visual, inertial, and depth sensors, determines a map of the robot's environment while localizing or identifying the position/pose of the robot within that map. However, the energy costs of running SLAM on real-time, mobile platforms can be expensive, limiting battery life for these devices in the wild. Thus, it is important to find energy-efficient pipelines for SLAM that can still obtain good accuracy and performance to enable ubiquitous robotic navigation.

Previous research for energy-efficient SLAM has utilized techniques such as motion planning and dynamic power management [1]. Real-time SLAM systems leverage efficient feature detection and description, local tracking and mapping, and parallel thread computing for fast performance [2]. Most of these approaches have concentrated on increasing computational efficiency after receiving sensor data.

Monocular visual SLAM is an emerging algorithm which has reduced the number and types of sensors necessary to a single visual camera, and has shown good localization results [2, 3]. Advantages include being lightweight in the hardware, applicable for mobile cameras and embedded platforms with low size, weight and power (SWaP). However, not much research has looked at the energy costs of image sensing itself for visual SLAM. In particular, image sensor processing (ISP) pipelines which convert RAW images to JPG/PNG images are normally tuned for creating highly aesthetic and visually pleasing images. It is unknown if this processing is needed for machine vision algorithms such as SLAM, and what optimizations can improve energy-efficiency without sacrificing accuracy.

In this paper, we investigate the effectiveness of visual SLAM on RAW images, without ISP processing, at varying types and levels of quantization. Image sensor quantization accounts for 50% of image sensing energy in modern CMOS image sensors [4], yielding significant opportunities for energy-efficiency in the pipeline. Our specific contributions include: (1) comparing linear and logarithmic quantization of RAW images with respect to localization accuracy for visual SLAM, and (2) introducing a new gradient-based quantization algorithm which quantizes the image spatially at various bit levels that outperforms both linear and logarithmic quantization for feature-based visual SLAM algorithms. We validate these contributions by testing two state-of-the-art visual SLAM algorithms on seven video datasets. This, to the best of our knowledge, is the first study to explore visual SLAM performance on RAW and varying quantized images.

2. RELATED WORK

Simultaneous localization and mapping (SLAM) has been an active area of research for over 30 years [5, 6], with recent advances in monocular visual SLAM algorithms [2, 3, 7, 8, 9]. For energy-efficient SLAM, eSLAM achieves real-time performance on low-power platforms by optimizing feature extraction and matching, yielding $41 - 71\times$ energy improvements and $1.7 - 3\times$ frame rate speed up [10]. Further, hardware acceleration such as a FPGA-based ORB feature extractor for SLAM reduced the energy consumption by 83% and reduced the latency by 41% compared to an Intel i5 CPU [11]. We are primarily concerned with optimizing the image sensing energy prior to the visual SLAM algorithm, and our methods are complementary with these systems.

The image sensor processing (ISP) pipeline utilizes demosaicing, denoising, color transforms, white balancing, and

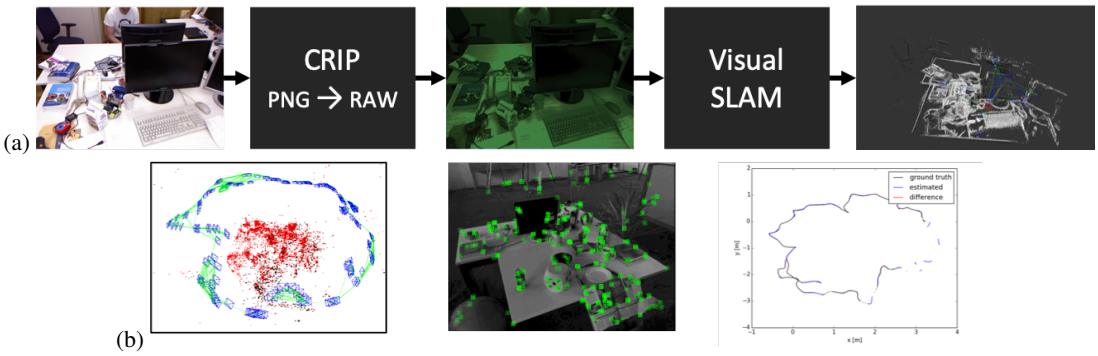


Fig. 1: (a) Experimental pipeline for analyzing quantization for Visual SLAM. The original dataset is run through CRIP to get RAW quantized images that are used for both SLAM methods. (b) ORB-SLAM2: Left - Scene mapping and camera trajectory. Center - Feature detection for a single video frame. Right - Output camera trajectory compared to the ground truth.

tone mapping to achieve high quality aesthetic images. However, for energy-efficiency, some smartphone cameras can bypass the ISP to produce RAW images. Liu et al. proposed an ISP that selectively disables stages depending on application needs [12]. The work most aligned with ours concerns reconfigurable ISP pipelines for energy-efficient computer vision [13]. This work shows how reduced ISP pipelines lead to vision accuracy-energy tradeoffs and save 75% of sensing energy with a minimalistic pipeline using logarithmic quantization. In our work, we leverage these insights and apply them to the particular case of visual SLAM. We introduce a new spatially-varying quantization method to improve the performance of visual SLAM over logarithmic quantization.

3. METHOD

Simulating RAW Data: The main challenge to evaluating the effectiveness of Visual SLAM on RAW data is the availability of suitably labeled datasets at varying quantization. We leverage the Configurable & Reversible Imaging Pipeline (CRIP) from [13] which can reverse JPEG/PNG images back to RAW format. CRIP was shown to have average pixel error of 1.064% and the PSNR was 28.81 dB as compared to real RAW images [13], lending confidence to the validation of our algorithms on this data. Using CRIP, we can convert visual SLAM data available online. For energy-efficiency, we turn off the ISP, including demosaicing, denoising, white balancing, color transforms, and tone mapping. This allows the sensor hardware to go straight from image sensor ADC to the SLAM algorithm, eliminating the ISP chip. Since visual SLAM, especially real-time systems, commonly work in grayscale intensity, most ISP optimizations are not critical.

Linear and Logarithmic Quantization. The image sensor analog-to-digital (ADC) converters operate on each pixel, and the typical linear ADC's energy cost is exponential in the number of bits in its output. Thus, image capture energy can be reduced via lower bit depths in sensor quantization, which can be achieved via successive-approximation (SAR) ADCs [14]. Quantization can be either linear or nonlinear. The nonlinear distribution of quantization levels can better represent images as the non-uniform probability distribution function for intensities in natural images is log-normal [15]. A central insight of Buckler et al. [13] was that log quantization uniformly mapped this distribution to equal bit values, thus performing approximate tone mapping of the images without the ISP. This yields beneficial accuracy/energy trade-offs across several computer vision benchmarks.

Gradient-based Quantization. In addition to linear and logarithmic quantization, we introduce a new form of quantization based on image gradients to help improve the accuracy-energy tradeoff. Our algorithm encodes regions with high-intensity gradient with higher bit values and lower gradient regions with lower values. Since most visual features contain gradient energy, this method preserves these features while downgrading non-salient regions at low bits. This yields significant energy savings in average bit depth across an image.

Gradient-based quantization relies on sensing the image gradient for pixels locally, and could theoretically be implemented in image sensor hardware. Focal-plane processing can compute basic functions such as edge detection and gradients in analog on the sensor [16], as well as optical pixels including Angle Sensitive Pixels [17] and event-based sensors [18]. While there is potential to implement this in hardware, for this study, this method is simulated using the pre-processed images for each quantization level.

Our algorithm is the following:

$$I_{GQ}[m,n] = I_{b[m,n]}[m,n], \tag{1}$$

$$b[m,n] = \min(\lceil \frac{W[m,n]}{\max(W[m,n])} * 7 \rceil + b_{min}, b_{max}). \tag{2}$$

where $W[m,n] = \sum_{(i,j)\in N(m,n)} \nabla I_{RAW}$ is the total gradient energy of a neighborhood $N(m,n)$ around pixel (m,n), ∇I_{RAW} is the image gradient magnitude, $b[m,n]$ is the bitmap which maps a pixel to a quantization bit depth, and $I_b[m,n]$ is the corresponding logarithmic quantized pixel at that bit depth. We use the gradient of the image using a 5×5 kernel. We use a 3×3 neighborhood, and shift all pixels

between 3 and 8 bits precision using $b_{min} = 3, b_{max} = 8$. In Figure 3, we show an example frame which has been quantized using our method. The red inlet shows an area where high gradient intensity is mapped to higher bit quantization, the green inlet surrounding the dice shows edge information, which is a mix of high and low quantization, and the blue inlet surrounding the floor with low gradient intensity is mapped to lower bit quantization.

Visual SLAM benchmarks. We deploy two benchmarks for Visual SLAM: ORB-SLAM2 [3], and LSD-SLAM [7], both of which are open-source real-time monocular SLAM systems. ORB-SLAM2 is a feature-based algorithm which detects features with ORB features, and then estimates the location and sparse depth map based on these features. ORB-SLAM2 performs four main tasks in parallel: tracking, mapping, re-localization, and loop closing. It is highly robust and compact via careful selection of only certain features and keyframes for reconstruction [3]. LSD-SLAM is a direct-based algorithm which utilizes the image intensities to estimate the location and semi-dense depth map. It is composed of three main parts: tracking, depth map estimation, and map optimization. The depth map is only created for pixels around large image intensity gradients [7].

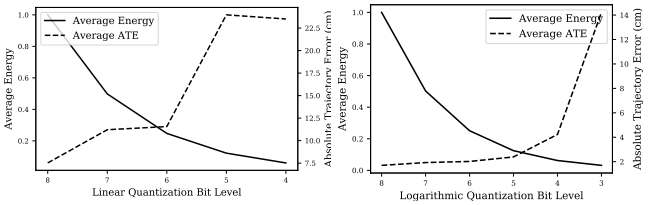
4. EXPERIMENTAL RESULTS

Dataset and Metrics. The TUM RGB-D benchmark dataset [19] was used to evaluate the accuracy of camera localization while running our imaging pipelines. This dataset provides sequences along with ground truth trajectory obtained with an external motion capture system. We utilize 7 videos from this dataset for our experiments, which although is smaller in scale, is on roughly the same order of videos evaluated as compared to the original ORB-SLAM [2].

Our error metric is absolute trajectory error (ATE) defined as the difference between points of the true and the estimated trajectory [19]. The true and estimated poses are matched via timestamps and then aligned using a similarity transform [20], as the scale of monocular SLAM is unknown. Then ATE is calculated as a root mean squared error.

While computational speed is another important metric, we do not report latency as we found that the per-frame processing time for all pipelines were roughly the same at 21-33ms on average. To quantify the expected energy savings of our imaging pipeline, we follow the model of [13] to compute the expected value of the ADC energy readout.

Initialization, Tracking and Features. Visual SLAM algorithms can suffer from issues with initializing at the beginning of the video, as well as maintaining tracking over the entire video. We observed that the number of features extracted while performing ORB-SLAM2 affects the performance of our quantization pipeline. As the bit level decreased in our quantization pipelines, the features were increased in order to preserve fast initialization and accurate tracking. We found that 4 bit linear and logarithmic quantization required an increase of four-hundred features, and logarithmic quantization lower than four bits required an increase of two-hundred features per bit level.



(a) Linear quantization (b) Logarithmic quantization

Fig. 2: Relative average ADC energy savings normalized to 8 bits and average ATE for ORB-SLAM2 over seven datasets with: (a) 4-8 bit linear quantization; (b) 3-8 bit logarithmic quantization.

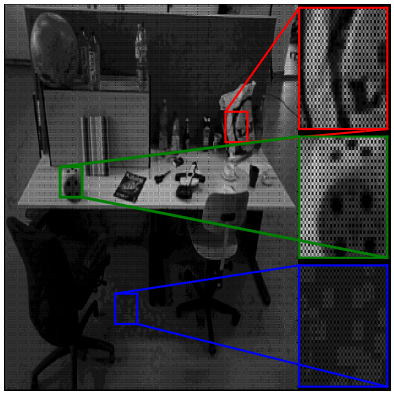


Fig. 3: Gradient-based quantization frame. The three inlets show differences in bit range with red (high gradient), green (mix of high and low), and blue (low gradient).

4.1. Analysis

Resolution. For ORB-SLAM2, we simulated resolutions $(640 \times 480, 533 \times 400, 427 \times 320, 320 \times 240)$ with ATE errors (0.29 cm, 3.48 cm, 6.46 cm, 10.9 cm) respectively for the fr2-xyz video. For the same video with the same resolutions, LSD-SLAM reported ATE errors (3.44 cm, 3.89 cm, 3.87 cm, 7.95 cm). We found similar trends for other datasets, but both SLAM algorithms failed to initialize and track below a resolution of 320×240. These experiments demonstrate that as the resolution decreases, visual SLAM accuracy degrades until it does not track for low resolutions. This means that image sensor subsampling such as windowing, ROIs, or binning would not be effective for these visual SLAM algorithms.

Frame Rate. We simulated frame rates of 30 FPS, 15 FPS, and 7.5 FPS by subsampling frames. For ORB-SLAM2 on two example videos, the ATE was constant down to 7.5 FPS. Below 7.5 FPS, however, it either failed to track or the error increased significantly. LSD-SLAM was sensitive to lower frame rates, with failure to initialize and track after 15

FPS. We observed similar trends in other videos. We hypothesize that low FPS causes feature matches to be more distant in time due to the frame subsampling, causing tracking issues.

Linear quantization: Linear quantization resulted in an increasing error trend with ORB-SLAM2, see Figure 2. The results show an average ATE of 7.52 cm at eight bits, 11.2 cm at seven and six bits, 4.23 cm at four bits, and 23.51 cm at the lowest working bit level of four bits. The average ATE was taken over four videos as two videos failed to initialize and track and one video resulted in very high error. None of the LSD-SLAM videos that were linearly quantized were able to initialize and track. To analyze this, it is helpful to look at the logarithmic quantization results for LSD-SLAM to draw comparisons.

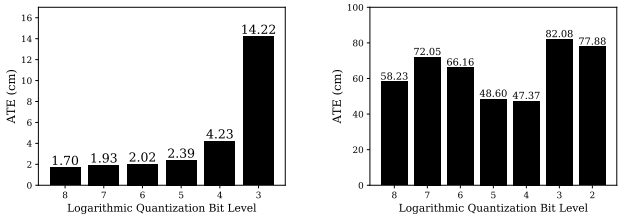
Logarithmic quantization: In Figure 4, we show the average ATE results over all datasets for our logarithmic quantization pipelines. ORB-SLAM2 was generally more robust to logarithmic quantization and shows an expected trend of increasing ATE as the bit value decreases, with a low ATE of 1.70 at 8 bits that increases to 4.23 at 4 bits, then jumps to 14.22 at 3 bits. Logarithmic quantization outperformed linear quantization because of the approximate tone mapping effect that occurs due to the statistics of pixel values in natural images (which was also observed in [13]).

The results for LSD-SLAM, shown in Figure 4b, are less consistent and in general, show poor performance for any RAW image pipeline. The minimum ATE achieved was 47.37 at 4 bits while the maximum ATE of 82.08 occurred at 3 bits. These averages are much higher for each pipeline than those measured for ORB-SLAM2. However, we note that log quantization still outperforms linear quantization, which failed to initialize.

We believe there are two mechanisms at play for the performance of LSD-SLAM. First, the approximate tone mapping of log quantization affects image intensities and contrast, and thus enables an intensity-based method like LSD-SLAM [7] to perform better with log quantization than linear quantization. However, even in the log quantized RAW images, the Bayer pattern likely causes false textures to appear in flat regions, and causes LSD-SLAM errors. We note that although we are operating with no ISP in this paper, we did try demosaicing on log quantized images and were able to achieve a more consistent performance for LSD-SLAM.

Gradient-based quantization: For ORB-SLAM2, our gradient-based quantization algorithm leads to further gains in energy efficiency. As shown in Table 1, the average bit value of each video is consistently between 4 and 5 bits with an overall average of 4.41 bits. Even with this relatively lower bit average of the images, visual SLAM achieves an average ATE of 1.81. This ATE is comparable to 7 and 8 bit logarithmic quantization pipelines with ATEs 1.93 and 1.70 respectively, saving effectively 3-4 bits in energy.

We note that the average bit level is low because a majority of pixels contain flat gradient information. Edges or high gradient pixels are a much sparser set of the total. Keeping only these high gradient pixels at higher bit ranges allows for features to still be detected easily while saving energy.



(a) ORB-SLAM2 (b) LSD-SLAM

Fig. 4: Average ATE for logarithmic quantization for ORB-SLAM2 and LSD-SLAM.

For LSD-SLAM, we do not see similar benefits for gradient-based quantization, like our previous quantization experiments. Since LSD-SLAM does not use features but rather intensity differences, including high bit regions in an otherwise low bit quantized image does not improve the performance as well as it does for feature-based methods.

	fr1 xyz	fr2 xyz	fr1 floor	fr1 desk	fr2 desk	fr3 long office	fr2 desk person	Avg.
Avg. bit	4.55	4.42	4.31	4.55	4.38	4.23	4.42	**4.41**
ATE (ORB)	1.04	0.28	2.9	1.8	0.82	4.9	0.94	**1.81**
ATE (LSD)	4.06	3.72	71.1	69.9	88.7	159.4	45.1	**63.1**

Table 1: Gradient-based Quantization Results

5. DISCUSSION

In this paper, we investigate visual SLAM on RAW images without ISP processing and varying sensor quantization. Our results indicate that for feature-based visual SLAM algorithms, namely ORB-SLAM2, using RAW images with logarithmic quantization at low bit levels can be energy-efficient and high performing. In particular, our novel gradient-based quantization algorithm achieved effectively 3-4 bits in energy savings without sacrificing performance. However, we note that our results on LSD-SLAM are not as conclusive since the intensity-based SLAM method does not rely on feature mapping. It remains as future work to try and adapt sensor quantization schemes that can benefit these direct-mapping methods. It would also be of interest to test our methods on a deep learning SLAM algorithm like DeepSLAM [21]. Also, there is an opportunity to optimize the SLAM algorithm itself for RAW data to extract the maximum performance while maximizing energy-efficiency.

Acknowledgements: This work was supported by NSF REU Site CCF-1659871 and ASU's Fulton Undergraduate Research Initiative (FURI) for O.C., NSF CCF-1909663 for J.R. and S.J., and the SenSIP Center at ASU.

6. REFERENCES

[1] Y. Mei, Y.-H. Lu, Y. Hu, and C. Lee, "A case study of mobile robot's energy consumption and conservation techniques," in *Proceedings of 12th International Conference on Advanced Robotics*, Jul. 2005.

[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[3] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[4] Y. Chae, J. Cheon, S. Lim, M. Kwon, K. Yoo, W. Jung, D.-H. Lee, S. Ham, and G. Han, "A 2.1 m pixels, 120 frame/s cmos image sensor with column-parallel $\delta\sigma$ adc architecture," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 236–247, 2010.

[5] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE Robotics & Automation Magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[6] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (slam): Part ii," *IEEE Robotics & Automation Magazine*, vol. 13, no. 3, pp. 108–117, 2006.

[7] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European Conference on Computer Vision*. Springer, 2014, pp. 834–849.

[8] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, 2007.

[9] F. Endres, J. Hess, J. Strum, D. Cremers, and W. Burgard, "3-d mapping with an rgb-d camera," vol. 30, no. 1, pp. 177–187, 2014.

[10] R. Liu, J. Yang, Y. Chen, and W. Zhao, "Eslam: An energy-efficient accelerator for real-time orb-slam on fpga platform," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.

[11] W. Fang, Y. Zhang, B. Yu, and S. Liu, "Fpga-based orb feature extraction for real-time visual slam," in *Proceedings of the IEEE International Conference on Field Programmable Technology*, 2017.

[12] Z. Liu, T. Park, H. Park, and N. S. Kim, "Ultra-low-power image signal processor for smart camera applications," *Electronics Letters*, vol. 51, no. 22, pp. 1778–1780, 2015.

[13] M. Buckler, S. Jayasuriya, and A. Sampson, "Reconfiguring the imaging pipeline for computer vision," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 975–984.

[14] R. J. Van de Plassche, *CMOS integrated analog-to-digital and digital-to-analog converters*. Springer Science & Business Media, 2013, vol. 742.

[15] W. A. Richards, "Lightness scale from image intensity distributions," *Applied Optics*, vol. 21, no. 14, pp. 2569–2582, 1982.

[16] M. A. Clapp, V. Gruev, and R. Etienne-Cummings, "Focal-plane analog image processing," in *CMOS imagers*. Springer, 2004, pp. 141–202.

[17] H. G. Chen, S. Jayasuriya, J. Yang, J. Stephen, S. Sivaramakrishnan, A. Veeraraghavan, and A. Molnar, "Asp vision: Optically computing the first layer of convolutional neural networks using angle sensitive pixels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 903–912.

[18] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128$\times$128 120 db 15μ s latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[19] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[20] B. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society of America A*, vol. 4, no. 2, pp. 629–642, 1987.

[21] R. Li, S. Wang, and D. Gu, "Deepslam: A robust monocular slam system with unsupervised deep learning," *IEEE Transactions on Industrial Electronics*, pp. 1–1, 2020.