Quantify the Effect of Histogram Intersection in Spatio-Temporal Data Sampling

CHANGFENG ZOU and JON C. CALHOUN (ADVISOR)*, Holcombe Department of Electrical and Computer Engineering - Clemson University, USA

The computational advance in high-performance computing leads to increased data generation by applications, resulting in a bottleneck within the system due to I/O limitations. One solution is the Spatio-temporal sampling method, which takes advantage of both spatial and temporal data reduction methods to produce higher post-reconstruction quality. Various user input parameters such as the number of bins or histogram intersection limit the performance for Spatio-temporal sampling. This poster focuses on determining the effect of the histogram intersection threshold in the Spatio-temporal sampling method. Results indicate that as long as a data set is not identical across adjacent time-steps, reducing the histogram intersection percentage increases the sampling bandwidth until blocks reused become static. The ExaAM data set shows an increase of 100-130% in sampling bandwidth, with only about a 5% decrease in PSNR value at 60% histogram intersection or lower.

Additional Key Words and Phrases: Histogram Threshold, Spatio-temporal Sampling, PSNR, Sampling Bandwidth, Time-step

1 INTRODUCTION

The improvement in computation speed of high-performance computing (HPC) increases the data generated by applications. Since the I/O performance is not on the same level as computation speed, resulting in bottlenecks within the system. This leads to the need for in situ data reduction to prevent bottlenecks from happening [2].

Two of the most popular data reduction approaches are compression and data sampling. Compression such as lossy compression is capable of achieving a high compression ratio through controlled error-bound [6]. Data sampling such as Spatio-temporal sampling reduces the data size by using both the spatial and temporal data reduction methods [4]. While this method is effective, it heavily relies on various user input parameters. This leads to lost performance when the user chooses a non-optimized setting. Fulp et al.'s Spatio-temporal sampling methods were all tested at 100% histogram intersection [4]. In this poster, we explore the impact of histogram intersection on the performance of the Spatio-temporal sampling from two aspects, sampling bandwidth and PSNR values.

This poster makes the following contributions:

- We depict the effect of histogram intersection percentage on sampling bandwidth and PSNR value on the Spatio-temporal sampling method.
- We determine that the effectiveness of histogram intersection depends on the similarity between different time-steps within the data set.

2 BACKGROUND

Spatio-temporal sampling takes a set of samples by leveraging both spatial statistics and information from neighboring time-steps (see Figure 1). The first step is to create a histogram representation of each time-step of the data set, with a user-defined number of histogram bins and pre-defined minimum and maximum range for the entire series [5]. Afterward, the data are sorted from least to greatest. Using the sorted values, the method develops an acceptance function that generates an acceptance rate between 0 and 1 based on the user-defined sample ratio. Then the data of the

Authors' address: Changfeng Zou, Changfz@g.clemson.edu; Jon C. Calhoun (Advisor), jonccal@clemson.edu, Holcombe Department of Electrical and Computer Engineering - Clemson University, 433 Calhoun Dr, Clemson, South Carolina, USA, 29634.

2022. Manuscript submitted to ACM

current time-step within the region is compared with the previous time-step to determine if a region is reusable based on the user-specified histogram intersection threshold. As long as the histogram intersection is above the histogram threshold, the method flags the region as reusable. To ensure no critical data is missed, each data value generates a random value between 0 and 1. If the value is below the acceptance rate and is not part of the previously-reused region flagged, then the value is flagged as a sampling stencil. Afterward, the method appends the sampling stencil and the chosen samples to the sample data array. As the Spatio-temporal sampling may not use the entire sampling budget, extra samples are gathered at the time step to ensure the highest post-reconstruction quality.

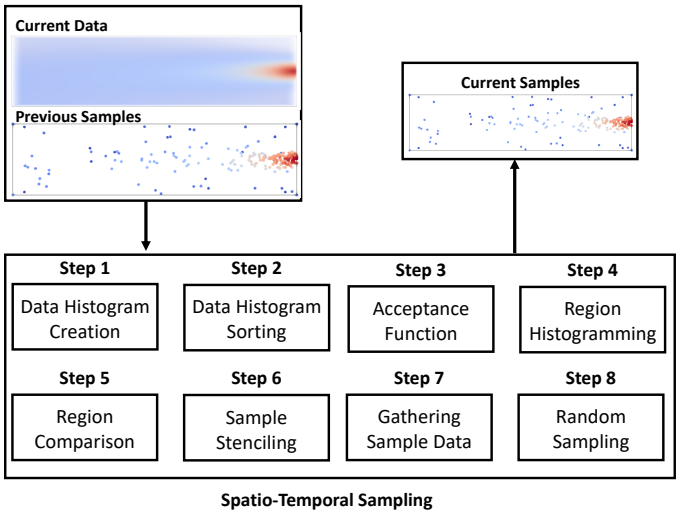


Fig. 1. Overview for the Spatio-Temporal Sampling Method [5]

3 METHODS

In order to ensure we are only testing the effect of histogram intersection, all the other configurations were the same one Fulp et al. used [4]. We begin by setting the 100% histogram intersection as the base case. Then compare the base case results with 0%, 50%, and 90% histogram intersection to determine the succeeding testing percentage until the percent of blocks reused becomes static.

4 EXPERIMENTAL RESULTS

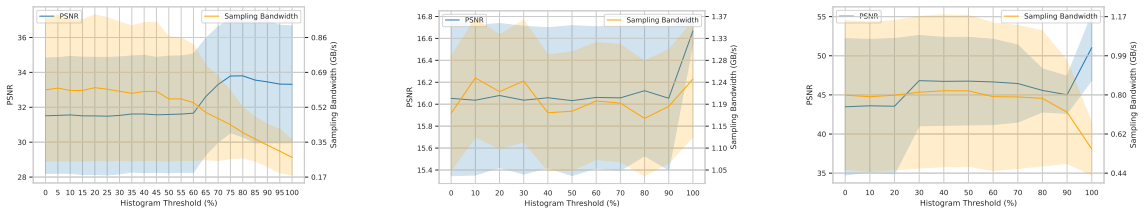
All the experiments are run on Clemson University's Palmetto Cluster on an Intel Xeon 6148G CPU and an NVIDIA v100 GPU. Each run uses a histogram-based reuse sampling method on GPU, nearest neighbors reconstruction, and a 1 percent sample ratio with CUDA version 11. For dataset-specific configuration is shown in Table 1. Figure 2 shows the overall changes in PSNR value and sampling bandwidth across histogram intersections for all the datasets tested in this experiment. The shaded region is the standard deviation, and the line is the average value.

4.1 ExaAM Dataset

The results from the ExaAM dataset in Figure 2a shows an increase of 100-130% in sampling bandwidth for histogram intersections lower than 60%, while the decrease in PSNR is within 5 percent compared to 100% histogram intersection.

Table 1. Configuration for the data sets used in this experiment

DATASET	DIMENSION	STEPS	REGION DIMENSION	BINS	INPUT MIN	INPUT MAX
ExaAM [1]	20 x 200 x 50	107	10 x 40 x 10	633	300.271	927.426
Asteroid Impact [7]	300 x 300 x 300	30	50 x 50 x 50	27	0	1
Hurricane Isabel [3]	500 x 500 x 100	10	25 x 25 x 25	27	-5471.85791	3225.4257



(a) ExaAM dataset (b) Asteroid Impact dataset (c) Hurricane dataset

Fig. 2. PSNR & Sampling Bandwidth at Different Histogram Threshold

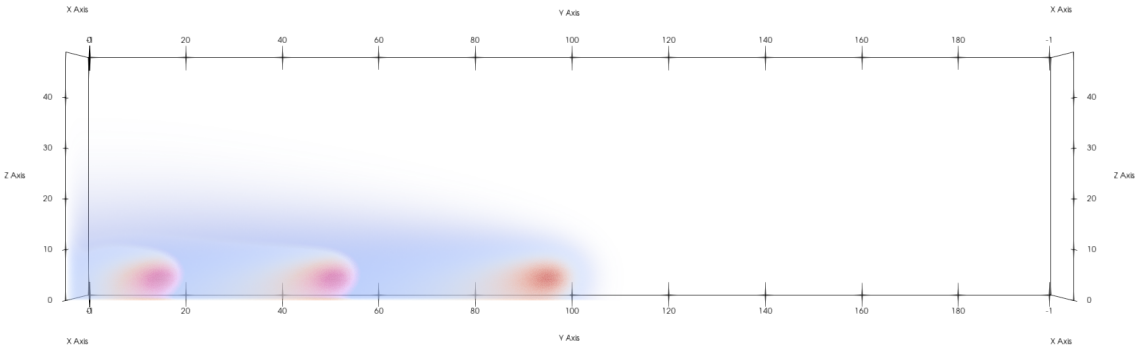


Fig. 3. Movement of ExaAM Data Across Selected Time-Step

This indicates that as the histogram intersection decreases, more regions are reused, which increases the sampling speed and decreases the PSNR as predicted. In Figure 3 we see the data moving across in space, and this fast movement of data indicates the region's changes at each time step. This results in an improvement in bandwidth as the histogram threshold decreases until the point where the percent of blocks reused becomes stabilized.

4.2 Asteroid Impact and Hurricane Dataset

In the Asteroid Impact data set, the result shows no improvement in PSNR or sampling bandwidth as the histogram intersection changes. This result is rather strange, since the predicted trend was that blocks reuse and sampling bandwidth increase as the histogram intersection decreases. Once we looked into the raw data for both datasets, we realized that, especially for the Asteroid Impact dataset, the time-steps are highly similar in that most regions remain the same. Even if we reduce the histogram threshold, it does not increase the number of blocks reused. Consequently,

changing histogram intersection does not affect the Asteroid dataset. On the other hand, since the Hurricane dataset has more movement across time-step, the sampling bandwidth increase as histogram intersection decreases.

5 CONCLUSION

The impact of histogram intersection on PSNR value and sampling bandwidth for the Spatio-temporal sampling method was not determined prior to this poster. Results indicate that improvement of sampling bandwidth is possible through histogram intersection as long as the dataset changes drastically throughout the time-step. For example, the ExaAM dataset doubles the sampling bandwidth as long as the histogram threshold is below 60%. Our approach begins with framing 100% histogram intersection as the base case and comparing the result with 0%, 50%, and 90% to determine the following succeeding testing percentage until the percent of blocks reused becomes static. Once we approach the static blocks reused region, that is also the region of most optimum histogram intersection.

ACKNOWLEDGMENTS

Clemson University is acknowledged for generous allotment of compute time on the Palmetto cluster. This material is based upon work supported by the National Science Foundation under Grant No. SHF-1910197 and SHF-1943114.

REFERENCES

[1] James Belak, John Turner, and ExaAM Team Team. 2019. ExaAM: Additive manufacturing process modeling at the fidelity of the microstructure. In *APS March Meeting Abstracts (APS Meeting Abstracts, Vol. 2019)*. Article C22.010, C22.010 pages.

[2] Ayan Biswas, Soumya Dutta, Earl Lawrence, John Patchett, Jon C. Calhoun, and James Ahrens. 2021. Probabilistic Data-Driven Sampling via Multi-Criteria Importance Analysis. *IEEE Transactions on Visualization and Computer Graphics* 27, 12 (2021), 4439–4454. https://doi.org/10.1109/TVCG.2020.3006426

[3] Hurricane ISABEL Simulation Data. 2019. http://vis.computer.org/vis2004contest/data.html online.

[4] Megan Hickman Fulp, Ayan Biswas, and Jon C. Calhoun. 2020. Combining Spatial and Temporal Properties for Improvements in Data Reduction. In *2020 IEEE International Conference on Big Data (Big Data)*. 2654–2663. https://doi.org/10.1109/BigData50022.2020.9378457

[5] Megan Hickman Fulp, Dakota Fulp, Ayan Biswas, Meilssa C. Smith, and Jon C. Calhoun. 2022. Accelerated Dynamic Data Reduction Using Spatial and Temporal Properties. *The International Journal of High Performance Computing Applications* (2022). in submission.

[6] Xin Liang, Sheng Di, Dingwen Tao, Sihuan Li, Shaomeng Li, Hanqi Guo, Zizhong Chen, and Franck Cappello. 2018. Error-Controlled Lossy Compression Optimized for High Compression Ratios of Scientific Datasets. In *2018 IEEE International Conference on Big Data (Big Data)*. 438–447. https://doi.org/10.1109/BigData.2018.8622520

[7] John M. Patchett and Galen Ross Gisler. 2017. Deep Water Impact Ensemble Data set. Technical report, Los Alamos National Laboratory. https://datascience.dsscale.org/wp-content/uploads/2017/03/DeepWaterImpactEnsembleDataSet.pdf