

Three-Stage Lossless Compression Technique For Processing PMU Big Data In Modern Smart Grid

*LI Baojie**, *CHEN Shuo[†]*, *LU Yuxin[†]*, *HE Weisheng[†]*, *ZENG Xiaoming[†]*

** IRCCyN, Ecole Centrale de Nantes, Nantes, France, 45300*

*[†] State Key Laboratory of Electrical Insulation for Power Equipment,
Xi'an Jiaotong University, Xi'an, Shaanxi Province, China, 710049*

** libaojie@stu.xjtu.edu.cn*

Keywords: Data compression; Lossless compression; Big data; PMU; Adaptive Huffman; Smart grid.

Abstract

With the widespread proliferation and application of high data sampling rate device- Phasor Measurement Units (PMUs) in modern smart grid, large volumes of data are produced in the wide-area monitoring system, which results in increasing challenge in real-time analysis, data storage and data transmission. This paper proposes an improved lossless compression technique which is fully based on the analysis of PMU data structure and consists by three stages algorithm for handling massive data. In the first stage, adjacent coordinate data of one period are operated subtraction and the waveform difference method is adopted to simplify data of other periods. Then, adaptive dynamic Huffman algorithm is selected to compress the processed data. Finally, Run Length Encoding method is employed to decrease furthermore the redundancy. The performance of the improved approach was evaluated under simulation using field PMU data. Experimental results under both dynamic and steady states presented at the end of the paper prove the efficiency of the proposed technique.

1 Introduction

In modern power grid, with the rapid development and placement of PMU equipment in wide-area monitoring system (WAMS), grid managers have become gradually conscious of the increasingly growing of PMU data and therefore strongly desire an efficient data compression method to process such volumes of data as to ensure the reliability of the power system. For instance, more than 584 PMUs have been installed in Western Electricity Coordination Council (WECC) until April, 2015^[1]. In China, the amount of PMU is reported to reach 2000 by 2014, specifically speaking, a regional power company with 100 PMUs will generate more than 50GB data each day with sampling rate at 30Hz, and over 120GB with higher rate of 100Hz which is widely applied in China power grid^[2]. Nevertheless, conventional grid monitoring and analytic systems are not designed to handle such volume big data and information explosion with great complexity.

Recent years, research dedicated on the PMU data compression emerge in large numbers. [3] presents a method using wavelet packet transform to reduce the data redundancy, but the compression result is not satisfying, which is the same obstacle faced by method of *zip* algorithm proposed by [4]. In [5], authors target to employ embedded zerotree wavelet technique, however the method is demonstrated to be unfit for processing real PMU data. Actually, most approach adopted by researchers bring straight the traditional techniques created for treating image or video data to treat PMU data, but ignore the adaptability when lacking understanding of PMU characteristic. In order to improve the actionability and efficiency of PMU data compression, this paper propose an improved a three-stage lossless compression and processing technique which is fully based on the analysis of PMU data structure and consists by three stages algorithm.

The remainder of the paper is organized as follows. Section II introduces briefly an overview of PMU data compression techniques. Section III presents the methodology of the proposed compression approach and the algorithm structure of three stages. Section IV evaluates the performance of data compression using field PMU data. Section V concludes the paper and points out some key points of further research.

2 PMU Data Compression

This section mainly introduces the challenges faced by PMU data compression in modern power grid and the two types compression methods.

2.1 Challenges of PMU data compression

When it comes to normal data compression, the compression operation is carried in the sending terminals or equipment to decrease the data redundancy so that the volume of data can be reduced for convenient data transmission in the communication systems; In the receiving terminals, the received and compressed data is decompressed on retaining original fidelity and looking for higher compression ratio (λ_{CR}) as denoted in (1).

$$\lambda_{CR} = \frac{\text{the number of raw data}}{\text{the number of compressed data}} \quad (1)$$

As for handling PMU data, additional challenges are attached to the compression assignment. In the first, the compression work is required to be executed on real-time, so as to realize the on-line monitoring of power grid condition and recognize low-frequency oscillations and other analysis of dynamic states. However, most of the current approach demand large sampling windows for the sake of better performance of compression, which therefore generate the difficulty of technique-application in real-time condition. Besides, regardless of the unexpected massive explosion of PMU information, the data compression work is supposed to be operated without sacrificing the integrity and accuracy of original PMU data in order to reserve critical grid information.

2.2 Types of data compression

According to the information theory, data compression is generally considered as the source encoding, which makes use of minimum data to transit maximum information. The basic process of compression is mainly based on the elimination of redundant information on keeping indeterminate ones by analysing the correlation and similarity of data source. Common compression methods can be divided into two categories- lossy and lossless types.

As a reversible compression method, lossless algorithm could reconstruct the original data without damaging the authenticity and complexity of information. Popular lossless approaches include arithmetic coding, run-length encoding, Huffman encoding and dictionary encoding, etc. Under the precondition of non-intervention of recognition and understanding, lossy compression methods seek for higher CR by sacrificing segmental data accuracy. Fourier algorithm and Wavelet encoding are the common lossy methods.

3 Methodology of three-stage compression

The block diagram of the proposed three-stage lossless compression algorithm is shown in Fig.1. After data acquisition from local WAMS/PMU collecting stations, the large volume of PMU data go through three stage of compression in succession, the methodology of which are discussed in Section 3.1, 3.2, 3.3, respectively. Afterwards, the communication channel transmits the well-compressed and real-time data to the processing centre for the subsequent data-reconstruction, which is the fundamental source for the advanced analysis and control of whole power grid.

3.1 Stage One: Adjacent coordinate and waveform difference methods

This stage targets to realize the pre-compression of primitive PMU information. Based on the communication theory, the algorithm adopted in this part takes the idea of Delta Modulation technique (DM) as reference to compare the instantaneous value of signal with quantized value captured at previous sampling moment, so as to utilize the correlative difference to preliminarily reduce the data redundancy.

- 1) *Adjacent coordinate difference method* is based on the correlation analysis of waveform data in one period when the amplitude of adjacent data varies continuously and show resemblance at some extent in one whole period during steady state.

$$U_a = \{U_{a1}, U_{a2}, U_{a3}, \dots\} \quad (2)$$

$$U_{ai} = [U_{ai}(1), U_{ai}(2), \dots, U_{ai}(x), \dots, U_{ai}(n)] \quad (3)$$

$$U'_{ai}(x) = U_{ai}(x) - U_{ai}(x-1) \quad (4)$$

The specific operating algorithm follows: first, separate the target data of voltage, phase and current of three-phase into corresponding periods as (2) (taking the voltage of phase-a for example); then, for each element of one period (3), replace the value $U_{ai}(x)$ by the difference between $U_{ai}(x)$ and $U_{ai}(x-1)$ as shown in (4), repeat the process until the first element, keep the value of $U_{ai}(1)$.

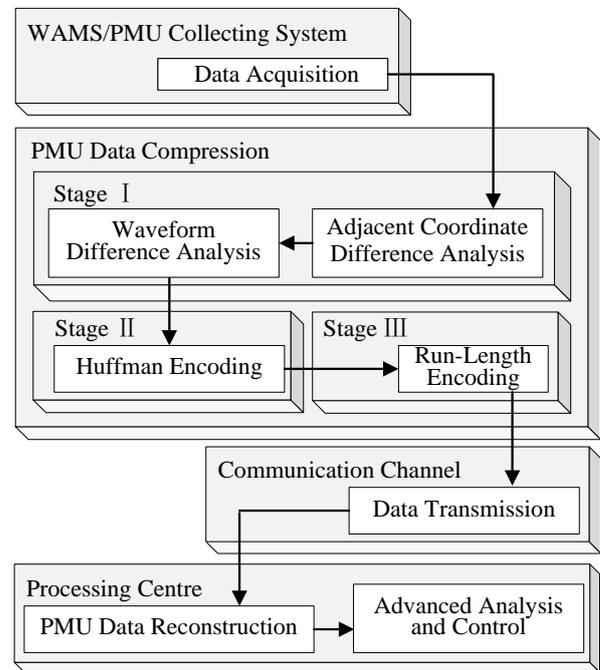


Figure 1. Block diagram of three-stage PMU compression

- 2) *Waveform difference method* utilizes the identical idea as adjacent coordinate difference method, but it compares the data among different periods on considering the similarity of corresponding phase under continuous and steady state. Therefore, the principle functions by: select one period as fundamental periodic sequence like U'_{a1} , make the value difference between each element in selected period with other period from all three-phase data as shown in (5).

$$U''_{ai}(x) = U'_{ai}(x) - U'_{a1}(x),$$

$$U''_{bi}(x) = U'_{bi}(x) - U'_{a1}(x), \quad x = 1, 2, \dots, n \quad (5)$$

$$U''_{ci}(x) = U'_{ci}(x) - U'_{a1}(x),$$

Expand the same procedure to the sequence of current and phase. If the value processed becomes negative,

replace it by the complement number, as to facilitate the compression work of subsequent stage.

3.2 Stage Two: Adaptive dynamic Huffman encoding

Adaptive dynamic Huffman encoding method could realize the construction and adjustment of Huffman tree at the same time, besides the efficiency of data transmission could also be ameliorated. During the encoding process, the ordering of node is required to keep to the following two properties described as the Sibling Property (SP):

- Node with higher weight (node attribute denoted to the occurrence frequency) has higher node number;
- Number of father node always exceeds the one of son.

In the course of weight-adjustment, necessary operation like exchange nodes should be executed in order to strictly ensure the SP. Due to the lack of the occurrence frequency of each node, the initial condition of encoding tree includes a symbol denoted as *NYT*, the weight of which is assigned as 0.

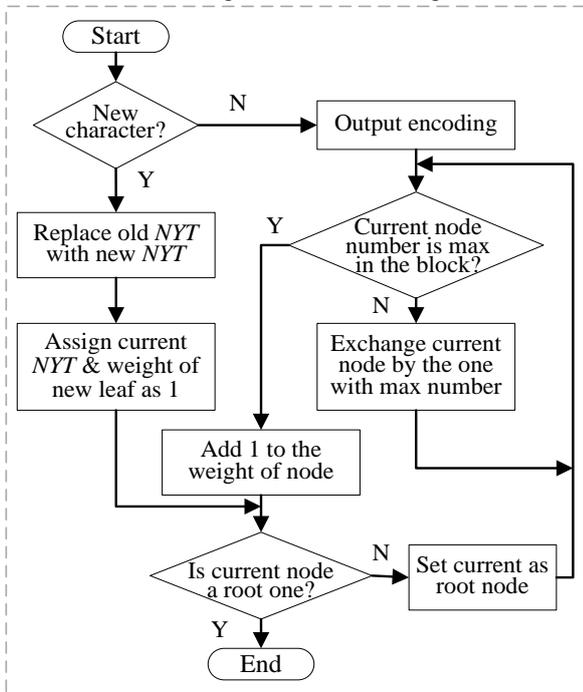


Figure 2. Flowchart of adaptive dynamic Huffman Encoding

When a new character occurs, a subtree with *NYT* and 2 leaf nodes needs to be constructed to replace the old *NYT* node, meanwhile the weight of the new character is assigned as 1. Each time the encoding of new character is accomplished, the corresponding weight increases by 1. The whole procedure of this adaptive dynamic Huffman encoding method is demonstrated in Fig.2 in detail.

3.3 Stage Three: Run-length encoding

Considering that the data compressed after Huffman encoding in stage two still contain a series of 0 code, the Run-Length Encoding (RLE) is adopted to furthermore reduce the redundancy of PMU information.

RLE utilizes three bytes to express one string, first byte represents the indication byte S_c , second for the byte of continuous occurrence, and the last for the repeat times as illustrated in Fig.4. It is observed that only under condition (6), the RLE method shows the significance with the λ_{CR} as (7).

$$L_{RUN} > 3 \quad (6)$$

$$\lambda_{CR} = \frac{L_{RUN}}{3} \quad (7)$$

Consequently, it is necessary to judge the value of run-length L_{RUN} before adopting the RLE method; During the decoding process, it is also required to judge S_c byte before determine the nature of the follow-up byte.

Source data stream:

0	0	0	1	1	2	2	2	2	3
---	---	---	---	---	---	---	---	---	---

Data stream compressed after RLE:

S_c	0	3	1	1	S_c	2	4	3	
-------	---	---	---	---	-------	---	---	---	--

Figure 4. Model of RLE compression method

4 Performance Evaluation

The whole procedure of the proposed three-stage lossless protocol compression technique is compiled under Matlab software. In order to precisely assess the performance of the mentioned techniques, the compression ratio of physical memory λ_{CRPM} is introduced and expressed as (8) on considering the actual bytes occupied for data-storage.

$$\lambda_{CRPM} = \frac{\text{storage space of raw data}}{\text{storage space of compressed data}} \quad (8)$$

All the compression tasks were executed on a desktop PC with RAM of 64 GB and processor of Intel Xeon E5-2680 2.70Ghz under Windows 10 system on using field PMU data obtained from 3 monitoring substations equipped with PMU devices located in France with sampling rate of 50 frames per second. The time of record covers the whole day of 31st March, 2015, when the European grid suffered a severe oscillation rooted from the Turkish blackout. The field record files were stored in as CSV format with comma separating each parameter.

4.1 Performance during steady and dynamic states

A record with time length of 100s from 07:34:00 of Central European Time (CET) was selected to evaluate the performance of the proposed approach under steady state and dynamic state with notable oscillation of frequency. The entire period is divided into 4 intervals of 20s as shown in Fig.3, where section A, B present steady state, C for dynamic, D denotes approximate steady state with slight oscillation.

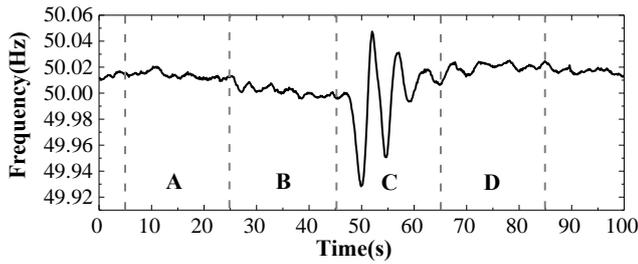


Figure 3. Field PMU record of frequency oscillation

The λ_{CRPM} results of three types of PMU data in float: voltage (amplitude and phase), current (amplitude and phase) and frequency during the four intervals are demonstrated in Fig.4. It is observed that the performance (average λ_{CRPM}) of the stages tested follows the order: A>B>D>C, which conforms to the oscillation level among the dynamic and steady states. Additionally, the λ_{CRPM} for current data is relatively lower than voltage and frequency data, which originates from the less accurate measurement of current and the interference of external noise that weakens the compression ratio of voltage data as well but less significantly.

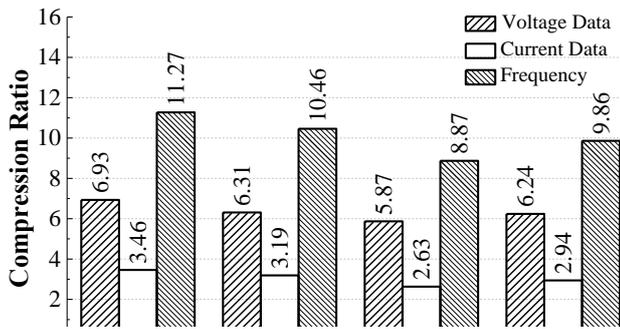


Figure 4. λ_{CRPM} of three types of data during four intervals

4.2 Performance comparison with other lossless means

For the sake of comparing the performance of other reported lossless algorithm, the *bzip2*^[6] based on the Burrow-Wheeler's block sorting, the *7-zip*^[7] based on the Lempel-Ziv-Markov algorithm, the *PPMd*^[8] algorithm based on non-dictionary prediction via partial matching together with the *gzip*^[4] which derives from the extended-rice algorithm are as well evaluated under identical circumstances. The comparison results are showed in Tab.1, in which the V_{mag} , V_{phase} , I_{mag} , I_{phase} and f represent the magnitude & phase of voltage, the magnitude & phase of current and the frequency, respectively.

Data type	Proposed method	<i>bzip2</i>	<i>7-zip</i>	<i>PPMd</i>	<i>gzip</i>
V_{mag}	5.36	4.36	4.58	3.61	4.71
V_{phase}	7.41	6.25	6.41	2.67	6.80
I_{mag}	2.65	1.98	2.03	0.86	2.34
I_{phase}	3.78	2.63	2.97	1.49	3.26
f	12.01	8.65	10.13	6.73	11.24

Table 1: Comparison results of PMU compression approaches

It could be concluded from the table that the proposed 3-stage lossless compression method performs well overall than other traditional approaches. Besides, the *bzip2* algorithm could also realize almost identical compression performance.

5 Conclusion

This paper develops a three-stage lossless compression technique for massive PMU data compression on overcome the shortcoming of conventional lossless compression approaches. Most of them adopt straight the algorithm created for treating image or video data to process PMU data, but ignore the adaptability when lacking understanding of PMU characteristic, consequently fail to achieve satisfying effect. On considering the structure of PMU data, the proposed method adopts three-stage algorithm to reduce the data redundancy step by step, including adjacent coordinate & waveform difference method, adaptive dynamic Huffman encoding algorithm and RLE technique, respectively. Through the evaluation of field PMU data, the presented method shows notable compression performance under both steady and dynamic states and displays superiority than other approaches reported in literature. Future research will be dedicated to more field experiments with larger volumes of PMU data and the issue of realizing real-time analysis.

Acknowledgements

This research work and the field PMU data were supported by the Mario Marceau and Andrea Peers at Réseau de Transport d'Electricité (RTE), France. The authors would like to sincerely acknowledge their assistance.

References

- [1] P. H. Gadde, M. Biswal, S. Brahma and H. Cao. "Efficient Compression of PMU Data in WAMS," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2406-2413, (2016).
- [2] F. Zhang, L. Cheng, X. Li, Y. Sun, W. Gao and W. Zhao. "Application of a Real-Time Data Compression and Adapted Protocol Technique for WAMS," *IEEE Transactions on Power Systems*, vol. 30, no. 2, pp. 653-662, (2015).
- [3] J. Khan, S. Bhuiyan, G. Murphy and J. Williams. "PMU data analysis in smart grid using WPD," *IECON 2014 - 40th Annual Conference of the IEEE Industrial Electronics Society*, pp. 2438-2444, (2014).
- [4] P. Top and J. Breneman. "Compressing Phasor Measurement data," *2013 IEEE Power & Energy Society General Meeting (PES)*, pp. 1-4, (2013).
- [5] J. Khan, S. M. A. Bhuiyan, G. Murphy and M. Arline. "Embedded-Zerotree-Wavelet-Based Data Denoising and Compression for Smart Grid," *IEEE Transactions on Industry Applications*, vol. 51, no. 5, pp. 4190-4200, (2015).

- [6] M. Burrows, D. J. Wheeler. "A block-sorting lossless data compression algorithm", *Digital Equipment Corporation, Palo Alto, CA, Tech. Rep. SRC-RR-124*, (1994).
- [7] I. Pavlov. (Apr. 2015). 7-Zip File Archiver. [Online]. Available: <http://www.7-zip.org/>.
- [8] A. Granados, K. Koroutchev and F. d. B. Rodriguez. "Discovering Data Set Nature through Algorithmic Clustering Based on String Compression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 3, pp. 699-711, (2015).