FAST–MeerKAT and SKA Pathfinders Synergies
B. Peng, C. Carignan and M. Zhu, eds.

Progress on FAST Extragalactic Survey Data Reduction Pipeline

Bo Zhang,[1], 2 Jian Xiao,[3] Ce Yu,[4] Qi Luo,[4] Zhicheng Yang,[4] and Ming Zhu[1,2]

[1]*National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China;* zhangbo@nao.cas.cn;mz@nao.cas.cn

[2]
CAS Key Lab of FAST, Chinese Academy of Sciences, Beijing, China

[3]*School of Computer Software, Tianjin University, Tianjin, China*
xiaojian@tju.edu.cn

[4]*School of Computer Science and Technology, Tianjin University;*
yuce@tju.edu.cn

Abstract. The extragalactic HI survey is considered as an essential way to , and one of the principle goals of the Five hundred-meter Aperture Spherical radio Telescope (FAST). Here we discuss the current progress on the development of HI survey's data reduction pipeline, with the aim of handeling large amount of observational data in a more automatic manner. And in this paper, we also present a deep-learning based algorithm for radio frequency interference flagging, as well as a CPU-GPU hybrid gridding code for data cube generation.

1. Introduction

The 21 cm hydrogen line (HI line) is considered as one of the most important spectral lines in the radio band. Emitted by hyperfine transition in neutral hydrogen atoms, it was proved to be an essential tool to outline the distribution of HI regions, both in the Milky Way and other galaxies. Massive HI "blind" surveys were made possible by the advent of multi-beam receivers for large single-dish telescopes in the late 20th Century, and tens of thousands of extragalactic HI emitters were discovered since then (e.g., see Haynes et al. 2018).

The Five hundred-meter Aperture Spherical radio Telescope (FAST; Nan et al. 2011), located in Guizhou Province in Southwest China, has already put extragalactic HI survey as one of its primary scientific goals. With the active parabolic reflector with an effective aperture of $\sim$ 300 m, a 19-beam receiver covering the 1.05 – 1.45 GHz band built by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Australia, and a maximum zenith angle of $\sim$ 40°, which provides an observable sky area twice larger than Arecibo's, the FAST Telescope brings new opportunity to extragalactic HI research, especially for low-mass hydrogen disks, and higher redshift ($z \gtrsim 0.3$) HI emitters.

The main challenge faced by FAST HI survery data reduction is the large volume of data. The spectral line backend provides 6, 5536 frequency channels in full Stokes for each beam, producing far more data than previous surveys, including Arecibo's ALFALFA Survey (Giovanelli et al. 2005; Haynes et al. 2018), the HI Parkes All-Sky Survey (HIPASS; Barnes et al. 2001; Meyer et al. 2004), as well as the HI Jodrell All-Sky Survey (HIJASS; Lang et al. 2003). To handel such a dataset, a high performance data reduction pipe line with minimal manual interactions is needed. In this paper, we will present the basic designs of the pipeline in Section 2, the current status of the development work in Section 3, and finally the discussion and conclusion in Section 4.

1.1. The Basic Scheme of the Data Reduction Pipeline

Refer to Arecibo's ALFALFA Survey, the complete HI data reduction process can be divided into 3 major stages, including the initial flagging and calibration of raw data (Level 1), building data cubes and signal extraction (Level 2), and finally doing research based upon results from Level 2 (Level 3), as shown in Figure 1. We adopted this scheme as the basic framework of FAST extragalactic survey pipeline. Our main principles in designing each stage include iterative data reduction, minimizing user input, and increasing efficiency.

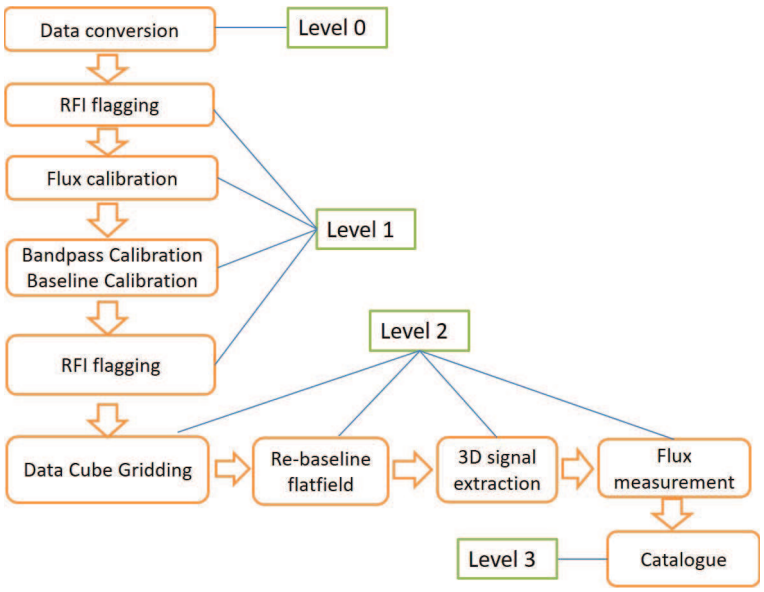

Figure 1. The flow chart of key steps for automatic HI data reduction, with all major stages noted.

Besides, to maximize our processing speed, as well as dealing with possible changes of format of raw data, we also propose a data conversion operation (Level 0). In this process, we will combine information related to telescope settings and operation logs with observational data, and convert the original FITS files into HDF5 files, for the convinience of parallel data reduction and storage.

2. Current Progress

Since the 19-beam receiver, the primary instrument of the FAST's sky survey projects, was just installed in May 2018, and is still under commissioning, currently our work mainly focus on later steps of data reduction, including radio frequency interference (RFI) identification and data cube preparation. In this section, we present our preliminary results on these issues. It should be noted that the current algorithms are for general use only. After the test runs are finished for the receiver, we will refine the current codes with observational data.

2.1. RFI Flagging with Deep Learning

In recent years, heuristic algorithms based on machine learning have been explored, in order to overcome the shortcomings of Traditional RFI mitigation methods using RFI physical characteristics (e.g., see Akeret et al. 2017a), including combinational thresholding methods and singular value decomposition (Offringa et al. 2010). Here we adopted a novel approach for automatic RFI recognition and flagging with an improved convolution neural network. The network was constructed based on U-net similar to Ronneberger's approach (Akeret et al. 2017), with a much deeper network structure to identify more components and complicated patterns, while minimizing errors caused by over-fitting.

Figure 2 shows the flagging performances of our network, as well as U-Net and SumThreshold, which is a variation of the combinational thresholding methods. The simulated data set, including astronomical signals (denoted by "TOD") and artificial interferences (denoted by "RFI"), is generated by the HIDE package (Akeret et al. 2017b). It serves as the standard reference (denoted by "RFI_Mask") of the flagging results. It is obvious that compared with other methods, our result provides the best approach to the reference, with nearly all RFIs correctly labeled.

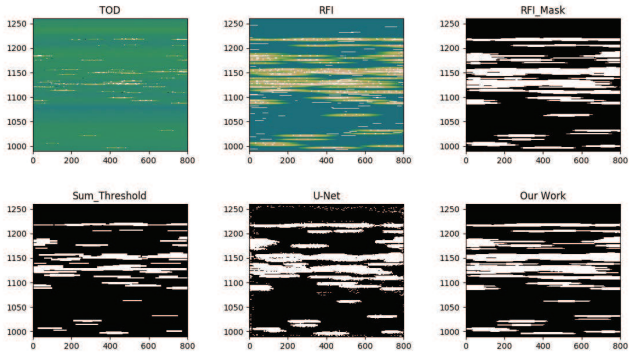

Figure 2. Comparison of RFI flagging results with various algorithms.

2.2. A CPU-GPU hybrid Convolution-based Gridding Algorithm

To generate uniformely-distributed data cubes with observational data, convolution-based gridding is the most commonly adopted approach. Basically speaking, the grid-

ding process carries out interpolation with weighting and averaging, rather than genuine convolution. A number of gridding packages have been developed, which can be divided into two major categories – scatter and gather, according to their ways of parallelization. Gridding via scattering iterates input data points, and calculates their contribution to a specified grid point (e.g., see Merry 2016), while gathering iterates output points to get an understanding of its possible contributors, and computes weighted average sum once and a[illegible]nd can provide a better resolution (e.g., see Winkel et al. 2016). Here we adopted the latter method, and developed a two-step algorithm based on CPU-GPU hybrid architecture to improve its performance.

The first step is to construct a pre-ordered lookup table based on HEALPix (Górski et al. 2005) to improve the efficiency of finding nearby contributors on CPUs. And secondly, we employ the high throughputs of GPUs, in order to accelerate the convolutional computation itself. Figure 3 shows the pre-ordering and searching scheme.

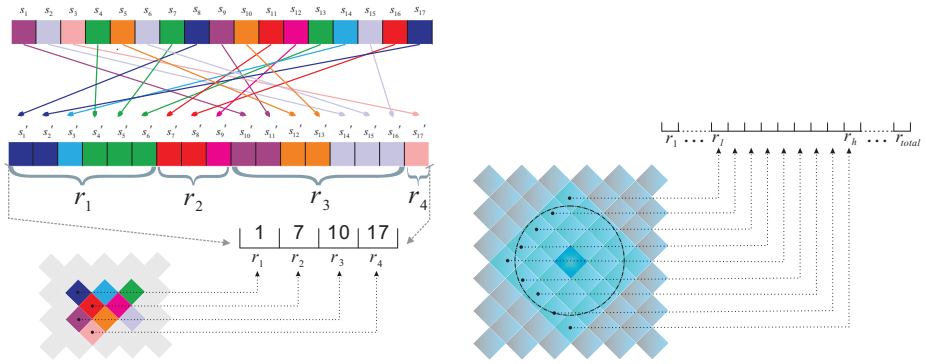

Figure 3. The pre-ordering (left) and searching (right) processes of gridding.

Figure 4 shows comparision of our result and Cygrid (Winkel et al. 2016), which is a high speed gridding implementations based on python and CPU multi-threading. Our method has much better performances on all input scales, especially for larger number of input data points. One reason for this is that, the linear compact pre-ordering is much faster than the hash-table-based preordering of Cygrid. Besides, as the volume of data increases, GPU's high throughput advantage becomes more significant.

3. Discussion and Conclusion

We proposed a data reduction pipeline prototype for FAST extragalactic HI survey. In order to deal with the large amount data produced by the FAST spectral line backend, we develop[illegible] novel algorithms to optimize the most time-consuming steps, including the RFI flagging and data cube gridding. We explored convolution neural network to identify RFIs automatically, with optimistic capability as shown by simulated data. We also developed a CPU-GPU hybrid convolution-based gridding algorithm, which can bring significant improvements on performance. These preliminary results showed a good prospect of the prototype, and we will refine the current works with more observational data in the near future.

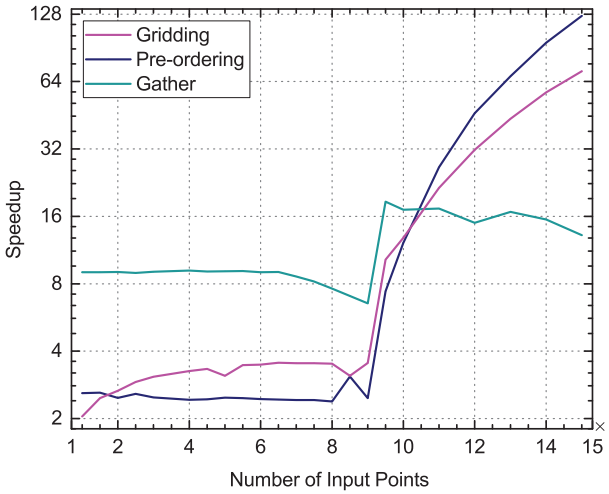

Figure 4. The performance of hybrid gridding approach compared with that of Cygrid.

Acknowledgments. The authors would like to thank Benjamin Winkel for providing the Cython code of the cygrid aglorithm.

The authors are supported by the Joint Research Fund in Astronomy (U1731125, U1531111) under a cooperative agreement between the National Natural Science Foundation of China (NSFC) and Chinese Academy of Sciences (CAS). BZ is also supported by the Young Researcher Grant of National Astronomical Observatories, Chinese Academy of Sciences.

References

Akeret, J., Chang, C., Lucchi, A., & Refregier, A. 2017, Astronomy and Computing, 18, 35

Akeret, J., Seehars, S., Chang, C., et al. 2017, Astronomy and Computing, 18, 8

Barnes, D. G., Staveley-Smith, L., de Blok, W. J. G., et al. 2001, MNRAS, 322, 486

Giovanelli, R., Haynes, M. P., Kent, B. R., et al. 2001, AJ, 130, 2598

Górski, K. M., Hivon, E., Banday, A. J., et al. 2001, ApJ, 622, 759

Haynes, M. P., Giovanelli, R., Kent, B. R., et al. 2018, ApJ, 861, 49

Lang, R. H., Boyce, P. J., Kilborn, V. A., et al. 2003, MNRAS, 342, 738

Merry, B. 2016, Astronomy and Computing, 16, 140

Meyer, M. J., Zwaan, M. A., Webster, R. L., et al. 2004, MNRAS, 350, 1195

Nan, R., Li, D., Jin, C., et al. 2011, International Journal of Modern Physics D, 20, 989

Offringa, A. R., de Bruyn, A. G., Biehl, M., et al. 2004, MNRAS, 405, 1150

Winkel, B., Lenz, D., & Flöer, L. 2016, A&A, 591, A12