Efficient Lossless Compression of Integer Astronomical Data

Òscar Maireles-González, Joan Bartrina-Rapesta, Miguel Hernández-Cabronero, and Joan Serra-Sagristà
Department of Information and Communications Engineering, Universitat Autònoma de Barcelona, Bellaterra, E-08193 Spain; oscar.maireles@uab.cat.

Abstract

Each new generation of telescope produces increasingly larger astronomical data volumes, which are expected to reach the order of exabytes in the next decade. Effective and fast data compression methods are paramount to help the scientific community contain storage costs and improve transmission times. Astronomical data differs significantly from natural and Earth-observation images, asking for specifically tailored compression approaches. This paper presents a novel lossless compression technique that employs the discrete Haar wavelet transform within the JPEG 2000 standard. Its performance is compared to that of a comprehensive selection of compressors, including fpack, the most common technique in astronomical observatories, as well as other algorithms highly competitive for other types of data. Experiments are performed on a large data set of 16 bit integer images, produced by telescopes around the world and representative of a wide variety of astronomical scenarios. The proposed technique has two modes. The first mode outperforms all the other tested techniques in terms of compression performance. It surpasses the most competitive configuration of fpack by, respectively, 5.3% (about 0.3 bits per sample), having also 4.5% lower compression and decompression times. The second mode is the fastest among all tested techniques. Its compression and decompression times are 2.5 and 3.5 times faster than the fastest configuration of fpack, while also yielding a 2.4% better compression performance (0.15 bits per sample).

Unified Astronomy Thesaurus concepts: Astronomy data reduction (1861); Astronomy data analysis (1858); Astronomy software (1855)

1. Introduction

Large volumes of astronomical data are generated daily in observatories around the world. Produced data volumes are growing both because of the increasing number of deployed telescopes and the higher resolution of new telescopes (de Zeeuw et al. 2014; Kremer et al. 2017; Rosa 2020). Figure 1 depicts the amount of data acquired by different surveys during the last three decades, as well as a forecast for the next decade. Increments of almost two orders of magnitude have taken place since the year 2000, and are predicted to reach the order of exabytes by 2030. Generated data are typically stored for use in current and future scientific studies, which results in increasingly large storage costs. In some observatories—specially in those using radio telescopes—, the storage requirements to archive all produced data has become prohibitive and some files need to be deleted to make room for newly acquired data (Bue et al. 2014; Kopczynska, 2022). To achieve sustainable storage costs, maximize the scientific value of available data and minimize transmission times, efficient compression techniques are required. Due to the significant differences between astronomical data and other types of imagery such as natural images and other remote-sensing data, compression algorithms must be tailored for this particular scenario.

NASA developed fpack (Pence et al. 2011), a suite of compression and decompression tools to handle astronomical images. The fpack software supports the following compression techniques: Rice, Hcompress, Gzip and PLIO IRAF. It has been deployed in most observatories, including European Southern Observatory (2014), Las Cumbres Observatory (LCO; 2022), Isaac Newton Group of Telescopes (2022), among many others. Its popularity is based on the fact that it can process data in the Flexible Image Transport System (FITS) (Hanisch et al. 2001) data format, the most common standard for storing uncompressed astronomical data. One main feature of fpack is its ability to divide the image in tiles before compression, which allows independent access to different spatial regions (White et al. 2012).

The main contribution of this paper is a novel method for compressing astronomical 16 bit integer data. The proposed approach uses the reversible discrete Haar wavelet within Part 2 of the JPEG 2000 standard, which has not been previously tested on astronomical data. Our proposal can operate in two modes, one aimed at achieving the best coding performance and another focused on achieving the shortest compression and decompression times. A second contribution is the gathering
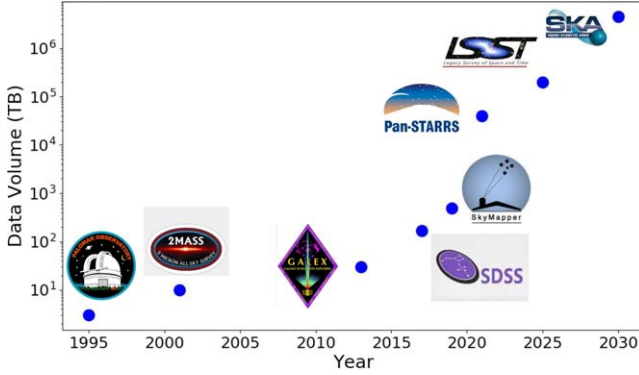
Figure 1. Evolution of Astronomical Surveys Data Volumes. Surveys correspond to, in chronological order, Palomar Digital Sky Survey (DPOSS); Two Micron All-Sky Survey (2MASS); Galaxy Evolution Explorer (GALEX); Sloan Digital Sky Survey (SDSS); Sky Mapper Southern Sky Survey; Panoramic Survey Telescope and Rapid Response System (PanSTARRS); Large Synoptic Survey Telescope (LSST); and Square Kilometre Array (SKA).

and publication of a new test data set comprising real data from multiple telescopes. This corpus contains a larger amount of samples depicting a wider variety of astronomical objects than previously described corpora (Schindler et al. 2011; Kitaeff et al. 2015; Pata & Schindler 2015), and is intended to be representative of modern observatories. Only the work by Pence et al. (2009) reports results for a larger data set, including sky and calibration images. Their work assesses the performance of Rice, Hcompress and Gzip coding techniques, but lacks the assessment of some modern and better performing compression techniques. Empirical results employing the new corpus suggest that the proposed method outperforms all other tested techniques, including fpack.

The rest of this paper is organized as follows. Section 2 describes the characteristics of astronomical images and the gathered test corpus. Section 3 puts forward our proposed method after revising related work. In Section 4, experimental results are presented and analyzed. Conclusions and future work are discussed in Section 5.

2. Astronomical Data and Compiled Dataset

2.1. Astronomical Data

The acquisition of data with a telescope gives as a result a 16 bit unsigned integer image, where each pixel value is proportional to the number of photons received by a sensor cell. The resulting image is normally stored in FITS format, which contains the pixel values plus the observation conditions of the acquired data. Images usually share the following structure. There is a dominant background with low pixel values that exhibit small variations, partly caused by atmospheric conditions (Yan et al. 2012). The foreground contains high-valued pixels depicting the astronomical objects of interest (Pence et al. 2009; Lin et al. 2021). Usually,

foreground astronomical objects have a small area, but some of them can fill an important part of the image surface. Depicted objects are normally star fields (as shown in Figures 2(a) and (b)), near galaxies (Figures 2(c) and (d)), star clusters (Figure 2(e)) or galaxy clusters (Figure 2(f)). Low exposure time images, as in Figure 2(a), commonly exhibit short pixel value ranges and low entropy values. Longer exposure times, as in Figure 2(b), increase the background pixel values and the number of saturated pixels, generally increasing the image entropy. In Figures 2(c), (e) and (f), the amount of luminosity of the observed field grows, inducing a less uniform background. This wide variety of background values, caused by the objects and their halos, tends to increase the image entropy. This effect is also seen when pointing to a faint object having luminous stars in the field, as observed in Figures 2(b) and (d). In all cases, the structure of astronomical images differs significantly from other types of imagery and makes efficient lossless compression challenging.

2.2. Compiled Dataset

To evaluate the efficiency of different compression algorithms, several astronomical images have been gathered from telescopes of different diameters located in observatories around the world. A data set of 225 astronomical images and 3.1 GB of raw data has been compiled. To obtain a representative data set, a wide mixture of situations and objects is considered, e.g., objects such as stars (isolated, open clusters and globular clusters), galaxies (elliptical, spiral, irregular, face-on, edge-on, isolated and interacting) and nebulae.

Selected telescopes from Roque de los Muchachos observatory are the Isaac Newton Telescope (INT), the Jacobus Katpeyn Telescope (JKT) and the William Herschel Telescope (WHT), with diameters of 2.5 m, 1 m and 4.2 m, respectively. Las Cumbres Observatory (LCO) data, a network of robotic telescopes, are also considered. Data from 0.4 m and 2 m diameter telescopes from McDonald, Haleakala, Cerro Tololo and Siding Springs observatories are included. Finally, data from the Joan Oró Telescope (TJO) at Observatori Astronòmic del Montsec, with a diameter of 0.8 m, are added too.

The characteristics of the different telescope images are summarized in Table 1. All data used for this research are publicly available at the following repository: https://gici.uab.cat/GiciWebPage/datasets.php#astronomical.

3. Proposed Approach

3.1. Related Work

A representative selection of lossless compression techniques is hereafter considered. The most popular techniques in astronomical observatories are Rice coding (Rice et al. 1993) and Hcompress (White et al. 1992), which uses the reversible discrete Haar wavelet transform followed by quadtree coding.
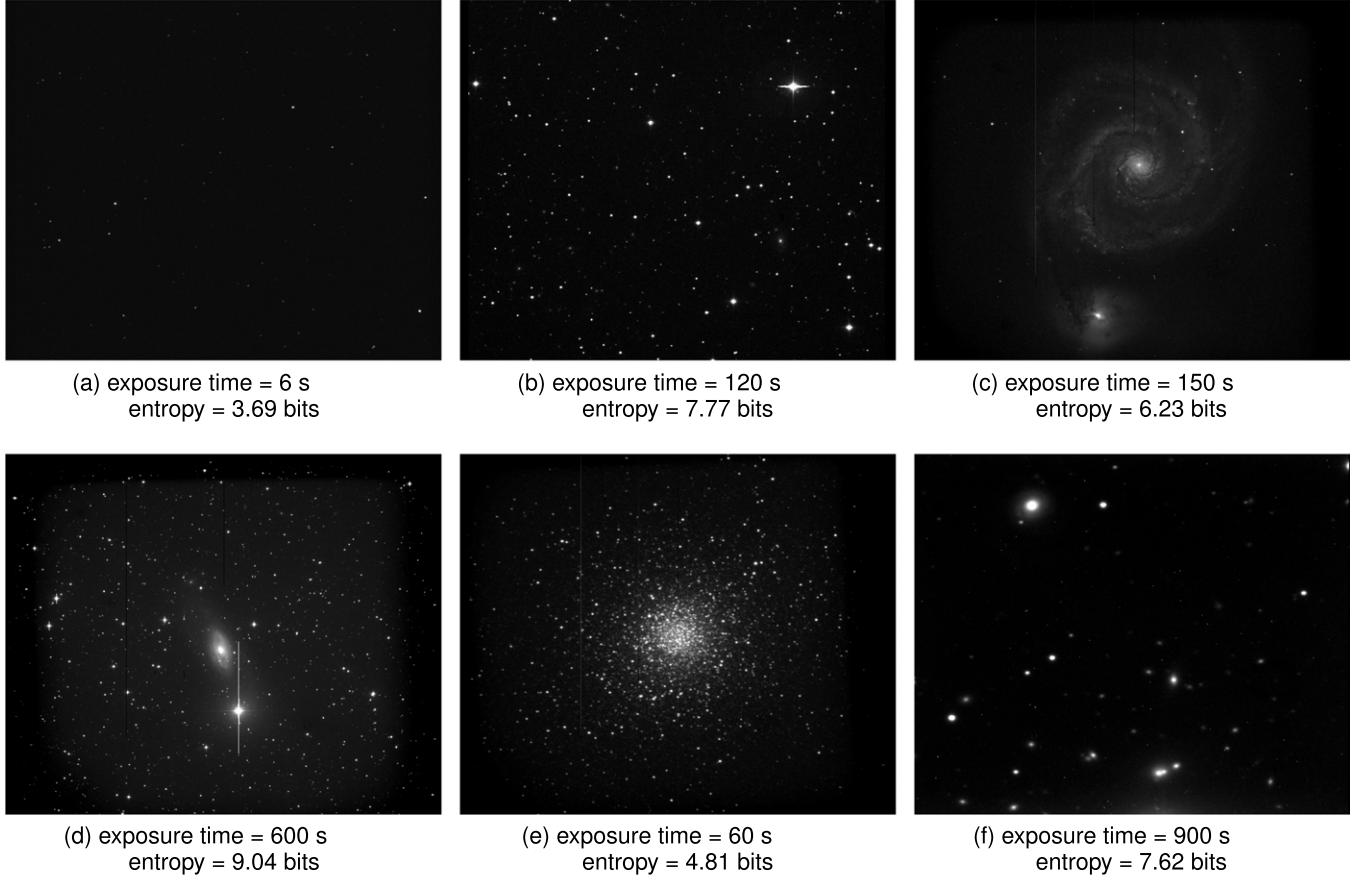
(a) exposure time = 6 s
entropy = 3.69 bits

(b) exposure time = 120 s
entropy = 7.77 bits

(c) exposure time = 150 s
entropy = 6.23 bits

(d) exposure time = 600 s
entropy = 9.04 bits

(e) exposure time = 60 s
entropy = 4.81 bits

(f) exposure time = 900 s
entropy = 7.62 bits

Figure 2. INT telescope images of (a) the NGC6791 field and (b) the M45 field. JKT telescope images of (c) the M51 field, (d) the NGC7013 field (e) the M5 field and (f) the Abell1066 field. Exposure time and entropy are shown below each image.

In a previous work (Maireles-González et al. 2022), JPEG 2000 (Taubman & Marcellin 2012), bzip2 (Seward et al. 2019) and JPEGLS (Weinberger et al. 2000) were found to achieve competitive compression performances. Lossless JPEG 2000 employs the 5/3 integer wavelet transform by default, followed by a bit-plane encoder and the MQ arithmetic coding (Medouakh & Baarir 2011). The bzip2 algorithm uses the Burrows–Wheeler transform, Run-Length Encoding and Huffman coding. Its highest compression level is selected. JPEGLS uses a predictor (with 3 causal neighbors), context modeling and a (Golomb 1966) entropy coder. Default parameters are used in the experiments.

Techniques not previously tested on astronomical data that employ different coding paradigms are also considered. The first paradigm is prediction-based coding. CCSDS 123.0-B-2 (Hernández-Cabronero et al. 2021) uses a predictor that can perform neighbor-oriented or column-oriented local sums followed by a sample, block adaptive or hybrid entropy encoder. CCSDS 123.0-B-2 results are obtained using its block-adaptive entropy encoder, reduced prediction mode and wide local sum (left, top, top-left and top-right neighbors), which yields the best results for astronomical data. HEVC

Table 1
Characteristics of Each Dataset Used

Telescope	Images	Dimensions	Average Entropy
INT	40	4200 × 2154	5.78
JKT	42	2100 × 2088	7.06
LCO	83	2112 × 3136	5.80
TJO	21	4096 × 4096	5.95
WHT	40	2501 × 2148	7.53

Note. Entropy results correspond to zero-order entropy.

(Sullivan et al. 2012) uses an intra prediction with directional modes and context-adaptive binary arithmetic coding. The HEVC predictor (Nair & Nair 2020) uses DC and Planar modes to encode the smooth image zones, e.g., the background of astronomical images. It also has 33 directional modes to predict directional or gradient structures. Default parameters for intra lossless compression are used to perform the experiments. FAPEC (de Mora et al. 2010) uses a linear predictor (Gimeno et al. 1994) and a custom prediction error coder. This predictor uses three causal neighbors (Ayoobkhan et al. 2017). The chunk size parameter is set to 8 MB and the adaptive block

Table 2
Entropy Values (in bps) Obtained by Weighting the Entropies of each Quadrant of a 9 Level Wavelet Transform

Dataset	Haar	5/3
ISO	5.291	**5.222**
Earth-observation	5.178	**5.055**
Astronomical	**5.594**	5.620

Note. Results for Astronomical and non-astronomical images are provided. Best results are highlighted in bold font.

length is set to 512 samples. The second paradigm is wavelet-based coding. FAPEC is also configured to use the integer 9/7-M discrete wavelet transform (DWT) instead of the aforementioned predictor (Hernández-Cabronero et al. 2020). The third paradigm is dictionary-based compression. LZMA (Ranganathan & Henriques 1993) uses a variant of the LZ77 (Ziv & Lempel 1977) algorithm with large dictionary sizes and an adaptive binary range coder. Zstandard (Collet 2017) applies the LZ77 algorithm followed by Finite State Entropy coding and Huffman coding. Both LZMA and Zstandard are configured to produce the best compression performance.

3.2. Proposed Technique

For most image types, the best wavelet transform for lossless compression tends to be the 5/3 filter bank (Taubman & Marcellin 2012). However, for astronomical data, this filter is not the best choice. Table 2 shows zero-order entropy results after applying 9 levels of the 5/3 and the Haar DWT filters when applied to natural, Earth-observation and astronomical images. Using 9 wavelet decomposition levels was empirically found to yield the best results. The natural image data set contains the ISO CCITT, ISO 12640-1 and ISO 12640-2 corpora. The Earth-observation data set contains the multi-spectral and hyper-spectral scenes employed by the Consultative Committee for Space Data Systems (CCSDS; Consultative Committee for Space Data Systems 2022) to design and evaluate CCSDS 123.0-B-2. Results shown in Table 2 reveal that the 5/3 DWT performs better for natural and Earth-observation images. However, the Haar DWT yields better results for astronomical data. To the best of our knowledge, this finding has not been previously disclosed for astronomical data.

Based on these results, a novel technique is proposed that employs the Haar DWT instead of the 5/3 DWT. To maximize coding performance for the test data set, 9 wavelet decomposition levels are employed. This transform is combined with the JPEG 2000 coding system, maintaining compliance with Part 2 of the standard (Lepley et al. 2002). This is the first mode of the proposed technique, focused on yielding the highest compression performance. The second mode aims at speeding up the compression and decompression times; it is referred to in the tables as "-HT." It complies with Part 15 of

JPEG 2000, also known as High Throughput (HT) JPEG 2000 (Taubman et al. 2019), which improves compression and decompression times in exchange for data reduction efficiency. All capabilities of JPEG 2000, such as scalability by position, resolution and quality are available in both modes. In contrast, fpack only allows tile-based spatial scalability. The additional features of JPEG 2000, as well as its maturity as an image compression standard, make both modes particularly adequate for archiving and dissemination purposes in astronomical observatories.

4. Experimental Results and Discussion

Lossless compression results are provided in terms of bits per sample (bps), lower values indicating better performance. Table 3 reports the average lossless compression results for all images of each telescope, and for all images in the corpus. Results are provided for all coding techniques described in Section 3.1, as well as the proposed technique described in Section 3.2. Table 3 shows that the first mode of the proposed coding technique provides the best compression results for each telescope, and for all images in the data set. On average for all images, it is 0.5 bps and 7.7% better than the most widely employed technique in astronomy, fpack with Rice. It is also about 0.3 bps and 5.3% better than the most efficient configuration of fpack, i.e., Hcompress. The Gzip and PLIO IRAF configurations of fpack perform worse than Rice and Hcompress, therefore their results are not included in this paper. The average image compression and decompression times are provided in seconds. The first mode of the proposed technique using 4 threads has, in addition, 4.5% lower compression and decompression times than Hcompress, as can be seen in Table 4. This first proposed mode is faster than 5/3 DWT JPEG 2000 due to its shorter filter length.

Techniques based on predictors show competitive although inferior results. The notable compression performance of JPEGLS is due to the high accuracy of the predictor in background regions and the efficiency of its context-adaptive encoder in regions depicting astronomical objects. In spite of the comparable results between JPEGLS and the proposed method, as well as its lower complexity, JPEGLS lacks desirable scalability and parallelization features present in fpack and JPEG 2000.

Tested techniques based on dictionary coding, i.e., LZMA and Zstandard, do not produce competitive coding results compared to prediction-based and wavelet-based methods. This is due to the fact that these techniques cannot directly exploit spatial redundancies present in images. A similar explanation justifies the low performance observed for the bzip2 and Rice algorithms.

The second mode of the proposed technique has the lowest compression and decompression times of all the tested techniques, as can be seen in Table 4. Since the other tested

Table 3
Compression Results in Terms of Average Bits per Sample

Telescope	DWT					Predictor			Dictionary		Others	
	Proposed	5/3 JPEG 2000	9/7 FAPEC	fpack-Hcompress	Proposed-HT	JPEGLS	CCSDS 123.0-B-2	HEVC	LZMA	Zstandard	bzip2	fpack-Rice
INT	**5.421**	5.436	5.531	5.775	5.750	5.479	5.540	5.555	5.714	6.147	5.561	5.949
JKT	**5.711**	5.734	5.817	6.055	6.050	5.758	5.824	5.891	6.284	6.954	5.986	6.209
LCO	**5.725**	5.757	5.811	5.994	6.061	5.754	5.743	5.828	5.944	6.414	5.850	6.144
TJO	**5.692**	5.723	5.756	5.909	6.008	5.778	5.735	5.850	5.979	6.404	5.856	6.121
WHT	**6.076**	6.101	6.186	6.513	6.391	6.170	6.176	6.485	6.728	7.425	6.348	6.619
All	**5.728**	5.754	5.824	6.051	6.057	5.782	5.798	5.910	6.109	6.646	5.913	6.204

Note. Best results in each row are highlighted in bold font.

Table 4
Average Image Compression and Decompression Times in Seconds

Technique	1 Thread Compression/ Decompression	4 Threads Compression/ Decompression
CCSDS 123.0-B-2	0.490 / 0.370	N/A
FAPEC	0.200 / 0.192	N/A
fpack-Hcompress	0.200 / 0.226	N/A
fpack-Rice	0.066 / 0.085	N/A
5/3 JPEG 2000	0.699 / 0.703	0.191 / 0.191
5/3 JPEG 2000-HT	0.043 / 0.049	0.028 / 0.030
JPEGLS	0.860 / 0.820	N/A
Proposed	0.705 / 0.712	0.190 / 0.189
Proposed-HT	0.042 / 0.048	**0.026 / 0.028**

Note. This average is calculated over all the data set. Best results are highlighted using bold font.

techniques have compression and decompression times higher than Rice and lack the scalability features offered by the proposed method and fpack, only the best compressors are included in this table. Only the implementation of JPEG 2000 supports multiple threads. Compression and decompression times of the second mode of the proposed technique using 4 threads are, respectively, 2.5 and 3.0 times lower than the fastest fpack configuration, i.e., Rice. Compression and decompression times are measured as the total user and system time reported by the operating system (Ubuntu 20.04 LTS) on a Intel Core i5 at 3.0 Ghz with 16 GB of DDR4 RAM. Execution times of the proposed technique can be further reduced with more threads and available central processing units. As shown in Table 3, this second mode produces average gains of 0.15 bps over Rice and equals Hcompress.

5. Conclusions

A competitive alternative to the compression technique commonly used in observatories, fpack, is proposed. The novel technique applies the reversible discrete Haar wavelet transform within the JPEG 2000 Part 2 standard. This wavelet transform exhibits better performance than the 5/3 DWT for astronomical data, contrary to what happens for other types of data such as natural and Earth-observation images. A large and representative astronomical data set has been gathered to evaluate the lossless coding performance of relevant compression algorithms based on different coding paradigms. Results indicate that the first mode of the proposed technique outperforms the compression performance of all other tested algorithms on average for each telescope and for all images in the corpus. It produces average gains of 0.5 bps and 7.7% over the most widely employed configuration of fpack, and 0.3 bps and 5.3% over its most efficient configuration. The second mode of the proposed method is the fastest of all the tested techniques. It has average image compression and decompression times 2.5 and 3.0 times lower than the fastest fpack technique, Rice. This mode produces average gains of 0.15 bps over Rice and equals the most competitive fpack compression technique, Hcompress. Other techniques provide competitive results, but lack the desirable scalability features offered by the proposed method and by fpack. Besides, the proposed technique includes other capabilities not present in fpack such as resolution and quality scalability for both modes. Parallelization is part of the algorithm out of the box and without any coding performance penalty. The improved coding efficiency and capabilities of the proposed method can help the scientific community reduce storage costs and transmission time, which is essential to deal with the fast-growing astronomical data volumes.

In the future, we aim at providing an implementation of this novel technique to the entire astronomical community through CFITSIO (Pence 2010). CFITSIO is an open source library of C and Fortran subroutines for reading and writing data files in FITS data format. CFITSIO supports multi-threading, enabling the implementation of our best performing approach. Having a CFITSIO implementation of our approach would allow to

integrate the compression technique into Fpack and Funpack software. We anticipate getting in touch with the FITS Support Office at NASA Goddard Space Flight Center to discuss this opportunity. Once the technique is implemented, compression results will be produced for the described 16 bit data set and also for 32 bit data.

Acknowledgments

References

Ayoobkhan, M. U. A., Chikkannan, E., & Ramakrishnan, K. 2017, EJASP, 2017, 1

Bue, B. D., Wagstaff, K. L., Rebbapragada, U. D., Thompson, D. R., & Tang, B. 2014, in Proc. 2014 Conf. on Big Data from Space (BiDS14), 206, ed. P. Marchetti & P. Soille (Luxembourg: Publications Office)

Collet, Y. 2017, Zstandard, v1.1.3, https://github.com/facebook/zstd

Consultative Committee for Space Data Systems 2022, Available: https://ccsds.org

de Mora, J. P., Villafranca, A. G., & García-Berro, E. 2010, JARS, 4, 041784

de Zeeuw, T., Tamai, R., & Liske, J. 2014, Msngr, 158, 3

European Southern Observatory 2014, Fpack, https://www.eso.org/sci/data-processing/faq/how-do-i-uncompress-fz-files.html

Gimeno, D., Torres, L., & Casas, J. R. 1994, in Proc. of 1st Int. Conf. on Image Processing (Piscataway: IEEE), 119

Golomb, S. 1966, ITIT, 12, 399

Hanisch, R. J., Farris, A., Greisen, E. W., et al. 2001, A&A, 376, 359

Hernández-Cabronero, M., Kiely, A. B., Klimesh, M., et al. 2021, IEEE Geoscience and Remote Sensing Magazine, 9, 102

Hernández-Cabronero, M., Portell, J., Blanes, I., & Serra-Sagristà, J. 2020, RemS, 12, 2955

Isaac Newton Group of Telescopes 2022, Archive, http://casu.ast.cam.ac.uk/casuadc/ingarch/

Kitaeff, V. V., Cannon, A., Wicenec, A., & Taubman, D. 2015, A&C, 12, 229

Kopczynska, Z. 2022, Mysterious source of fast radio bursts, Nicolas Copernicus University, https://portal.umk.pl/en/article/mysterious-source-of-fast-radio-bursts

Kremer, J., Stensbo-Smidt, K., Gieseke, F., Pedersen, K. S., & Igel, C. 2017, IISys, 32, 16

Las Cumbres Observatory (LCO) 2022, Fpack usage, https://lco.global/education/observing/fpack/

Lepley, M., Houchin, J. S., Kasner, J., & Marcellin, M. 2002, JPEG 2000 Extensions (Part 2), in The JPEG 2000 suite, ed. P. Schelkens et al. (New York et al.: Wiley), 71

Lin, B., Yang, X., Wang, J., et al. 2021, IGRSL, 19, 8012405

Maireles-González, Ò., Bartrina-Rapesta, J., Hernández-Cabronero, M., & Serra-Sagristà, J. 2022, in 2022 Data Compression Conference (DCC) (Piscataway, NJ: IEEE), 389

Medouakh, S., & Baarir, Z. 2011, IJCA, 18, 27

Nair, P. S., & Nair, M. S. 2020, Procedia Computer Science, 171, 1887

Pata, P., & Schindler, J. 2015, ExA, 39, 495

Pence, W., Seaman, R., & White, R. 2009, PASP, 121, 414

Pence, W., Seaman, R., & White, R. 2011, arXiv:1112.2671

Pence, W. D. 2010, CFITSIO, v2. 0: a new full-featured data interface, Astronomical Data Analysis Software and Systems VIII in ASP Conf. Ser. 172, ed. D. M. Mehringer, R. L. Plante, & D. A. Roberts, 487

Ranganathan, N., & Henriques, S. 1993, ITCS, 40, 96

Rice, R., Yeh, P.-S., & Miller, W. 1993, in 9th Computing in Aerospace Conf. (Reston, VA: AIAA), 4541

Rosa, R. R. 2020, AnABC, 93, 1

Schindler, J., Páta, P., Klíma, M., & Fliegel, K. 2011, AcPol, 51, 1

Seward, J. 2019, bzip2, v.1.0.8, http://www.bzip.org/

Sullivan, G. J., Ohm, J.-R., Han, W.-J., & Wiegand, T. 2012, ITCS, 22, 1649

Taubman, D., & Marcellin, M. 2012, Introduction to JPEG2000, JPEG2000 Image Compression Fundamentals, Standards and Practice, 642 (Berlin: Springer), 399

Taubman, D., Naman, A., Mathew, R., Smith, M., & Watanabe, O. 2019, International Telecommunications Union (ITU), 444, 15

Weinberger, M. J., Seroussi, G., & Sapiro, G. 2000, ITIP, 9, 1309

White, R., Postman, M., & Lattanzi, M. 1992, Compression of the Guide Star Digitised Schmidt Plates (Dordrecht: Kluwer), 167

White, R. L., Greenfield, P., Pence, W., Tody, D., & Seaman, R. 2012, arXiv:1201.1336

Yan, L., Jin, M., Fang, H., Liu, H., & Zhang, T. 2012, IGRSL, 9, 672

Ziv, J., & Lempel, A. 1977, ITIT, 23, 337