

Copyright 2002 Society of Photo-Optical Instrumentation Engineers. This paper will be published in Proc. Astronomical Telescopes and Instrumentation 2002 and is made available as an electronic preprint with permission of SPIE. One print or electronic copy may be made for personal use only. Systematic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication of any material in this paper for a fee or commercial purposes, or modification of the content of this paper are prohibited.

# Comparison of algorithms for the efficient approximation of heterogeneous multidimensional scientific data

Immanuel Freedman, Dr. Immanuel Freedman, Inc.

## ABSTRACT

Many scientists would like to be able to view and analyze quick look astronomical data on hand held devices linked by wireless network to the Internet. Scientific data is often characterized by high dynamic range together with abrupt, localized or extended changes of spatial and temporal statistical properties. I compare the effectiveness of algorithms for the efficient approximation of scientific data that support low bit-rate, near real-time and low-delay communication of heterogeneous multidimensional scientific data over existing or planned wireless networks.

**Keywords:** data analysis–image processing–astronomical data bases:miscellaneous

## 1. INTRODUCTION

Many scientists—including astronomers—are aware that hand held devices offer inexpensive mobile displays together with near-laptop computing and wireless communications capabilities. The recent advent of large virtual observatories (European AVO and NASA's Sky View...), orbiting telescopes (HST, NGST, Chandra...), manned space missions (Mission to Planet Mars...) and large ground-based CCD spectrographs (SUBARU...) all indicate an urgent need for the development of improved quick look imagery that can be displayed for rapid decision-making on mobile hand-held devices linked by wireless network to the Internet.

In this paper, I propose solutions that highlight the role of approximation in the visual, symbolic and graphic presentation of astronomical and engineering data for intuitive understanding and exploratory data analysis. Accurate astrometry and photometry may, in principle, be made available on demand for user-selected celestial objects, regions of interest, full sky background, ancillary measurements or engineering data sets.

The horizon time for taking action based on decisions made with the help of quick look data may vary widely with application. For example, the ALADIN project<sup>1</sup> considers that 20 seconds is the maximum delay an end-user can be expected to wait for a 2MB archived image served over the Web. The SUBARU telescope Quicklook Producer<sup>2</sup> is run daily at the Hilo Base Facility and typically generates about 500 kB compressed imagery from a 16x10MB data set based on a 10-15 minute observation made by the Suprime-Cam and other instruments. Such quick look data could take up to 60-200 seconds to transfer over a 14.4-28.8 kbps phone-grade network connected to the Internet. If a network of observers were to be provided with quick look data served to inexpensive mobile hand held platforms, they could place telephone calls or initiate remote system commands within minutes of discovering serious instrument or telescope malfunctions (e.g., the collapse of the Green Bank telescope or the failure of a gyroscope on board the COBE mission) or claim priority within hours for the serendipitous discovery of astronomical objects.

In this paper, I adopt cost and performance goals for a system based on mobile hand held devices of about 25-50% of the capital cost and about 100% of the operational cost of a system based on laptop computers, together with performance and quality goals of retrieving and displaying pre-processed data of visual quality satisfactory

to the user within about 30 seconds of a user demand.

The determination of visual quality is application-dependent. The hand-drawn sketches with which Galileo announced the discoveries of Jovian satellites<sup>3</sup>, Saturnian rings and the existence of Sun spots are strikingly simple. In the analogous field of cartographic generalization<sup>4</sup>, map makers extract and simplify relevant information and present it at smaller scales in a less cluttered<sup>5</sup> way, while meeting defined specifications on topology or geometry. In this paper I focus primarily on subjective quality assessments with only brief discussion of computation time and encoding delay.

Astronomical data is often heterogeneous in that variables of differing statistical properties are often stored in the same file or data stream but are not necessarily optimally organized for data compression. Such data is often also multidimensional and samples many associated variables. Examples include hyperspectral imagery (e.g., COBE/DIRBE data, or SUBARU/HDS echelle spectrography), imagery that evolves with time (e.g., SOHO observations of solar activity) and astronomical catalog entries indicating observational variables including object name, epoch, coordinate system, position, spectral class, spectral filter, time, exposure and observer's comments together with derived physical quantities including distance, redshift, mass, temperature and velocity dispersion.

The plan of the paper is as follows. In Section 2, I present a system overview and compare hand held devices with laptop computers in terms of specification and cost. I also compare existing or planned wireless with broadband networks in terms of available bandwidth and data transmission cost. In Section 3, I develop novel methods and combinations of methods of data approximation. In Section 4, I compare algorithms in terms of visual quality, bit rate and encoding delay. In Section 5, I present a summary of the findings and develop conclusions.

## 2. SYSTEM OVERVIEW AND COST COMPARISON

Platform	Processing Resource	Operating System	Display	Expansion Options	Manufacturer's Suggested Retail Price
PocketPC	Intel Strong ARM	Win CE Linux?	240x320x16	SD Card Slot	about \$730
FOMA	Unknown	Unknown	176x144x12	Unknown	about \$500
Palm	Motorola 68K	PalmOS Linux?	160x160x8 (effectively 160x120x8)	SD Card Slot	about \$199
Laptop	x86	Many	1024x768x24	Many	about \$1k to \$5k

Table 1 Comparing the cost and specification of platform resources.

Network	Max (Typical) Speed	Suggested Fixed Cost	Suggested Variable Cost
UMTS	2 Mbps	Unknown	Unknown
FOMA	384 kbps (64 kbps now, 325 kbps next phase)	About \$30/month	About \$4.2E-4 per 128-byte packet
GPRS	128 kbps (40 kbps )	Unknown	Unknown
iMode	9.6 kbps (5 kbps )	About \$10-20/month	About \$2.5E-3 per 128-byte packet
xDSL 1.5 Mbps	1.5Mbps (300kbps)	About \$30/month	None

Table 2. Comparing the cost and specification of network resources.

Figure 1 indicates the primary components of a system based on hand held devices for the evaluation of quick look astronomical data. Although there are still unknowns in the cost and capability information, Table 1 and Table 2 clearly show that systems based on hand held devices (especially those based on FOMA) can be cost-competitive with laptop systems for astronomical quick look use since at best they can deliver almost 10 MB image data in about 30 seconds. A modest compression ratio of 16:1 would suffice to deliver SUBARU

observations comprising 16x10 MB in less than 30 seconds.

In passing, note that a hand held system based on the Linux operating system may become capable of running popular astronomical image processing packages such as ESO-MIDAS, AIPS or IRAF, IDL or MATLAB.

### 3. DATA APPROXIMATION ALGORITHMS

1.Acquisition	2*.Preprocess	3.Encode	4*.Postprocess	5.Transfer	6.Decode	7.Display
From source	None	See Table 4	None	Network	per Table 4	Dither
From Source	Unsharp Mask	See Table 4	Deconvolution	Network	per Table 4	Dither

Table 3. Description of the overall sequence of steps (including optional steps denoted by \*).

Method	1.Transform	2.Quantize Coefficients	4.Organize Coefficients	5 Reduce Geometrical Redundancy	6. Reduce Statistical Redundancy	7.* Model. Residuals
DCTune	DCT 8x8	Perceptual	Zigzag sequence	Runlength coding	Huffman	Rice coding, Summary statistics, Discrete Sine Transform
JPEG-2000	Wavelet	Perceptual	Pyramidal	Yes	Yes	Rice coding, Summary statistics
FRIT	Finite Radon + Symmlet (4)	Retain top 0.5%	Zigzag sequence	Index of non-zero coefficients	gzip	Rice coding, Summary statistics
Spectral decorrelation	PCA	Retain top 20% of components	Linear, spectral	Follow with spatial coding	Huffman or gzip	Rice coding, Summary statistics
Spectral block matching	MPEG-4	Perceptual	Zigzag sequence	Runlength coding	Huffman	Rice coding, Summary statistics
Entropy	gzip Arithmetic	No	Organized by data variable	None	LZW Arithmetic	None
DPCM	DPCM	No	Non-causal predictor	None	None	Rice coding
Connected-component labeling	Morphological	Yes	Linear	Yes	Yes	Summary statistics
Cartographic generalization	Gradient Vector Flow snake	Line simplification	Linear, possible tree	Yes	gzip	Summary statistics

Table 4. Description and comparison of the sequence of steps used in the Encode step of Table 3..

Table 3 and Table 4 indicate the sequence of steps and combinations of methods of data approximation I used to explore the relationship between visual quality, bit rate and encoder delay.

DCTune<sup>6</sup> provides perceptual quantization matrices for the Independent JPEG Group encoder .I set the viewing distance at 16 inches for hand held devices rather than the usual 23.5 inches for desktop computers. JPEG-2000<sup>7</sup> is a scalable wavelet-encoder based standardized by the JPEG committee. FRIT is an extension of a Finite Ridgelet Transform<sup>8</sup> encoder in which the retained coefficients are encoded initially by

position and subsequently entropy coded. Decorrelation<sup>9</sup> of the spectral channels provides an effective color space for subsequent spatial coding. MPEG-4 provides motion compensation by block matching for sequences of movie frames with strong spatial correlation. Although the spectral channels are certainly not temporal in nature, if strong spectral correlations exist, the same block matching concept may, in principle, be used to reduce spectral redundancy. DPCM extends ENCODE<sup>10</sup> to multiple dimensions through a non-causal nearest neighbor predictor. Cartographic generalization refers to a method of segmenting images by active contours known as Gradient Vector Flow snakes<sup>11</sup> for which methods of line simplification<sup>12</sup> and generalization are known.

The DPCM has the lowest encoder delay as only the first or second nearest neighbor pixels which constitute the Minimum Coded Unit (MCU) need buffering. The Spectral Decorrelation Method reduces the delay when compared to methods that process all spectral channels. DCTune and JPEG-2000 conventionally buffer a full frame of imagery; however the MCU is 8x8 pixels for each spectral channel to be processed. Likewise the MPEG-4 MCU may, in principle, be as small as a Group of Blocks (for H.263 compatibility) but is usually the size of a Group of Pictures (at least 2 pictures for P frames, 3 pictures for B frames). The minimum coded unit for the FRIT method is one block of pixels constituting a finite projective space. For this application I considered blocks whose size in each dimension is a power of prime numbers. The encoding delay for entropy coding depends on the application and the size of buffer. Note, in passing, that Arithmetic Entropy Coding has, in principle, no delay. Connected Component Labeling refers to a morphological technique for creating a graph sketch of extended objects which in turn can be visually presented as a sketch. The MCU is the size of the connected set used in the search process. The simplification and generalization techniques used in the Cartographic Generalization approach may have delays ranging from a line segment (to be processed by the Douglas-Peucker algorithm) to a full frame.

JPEG-2000 and DCTune (JPEG) blur images and JPEG displays blocking artifacts for compression factors exceeding about 40:1. I explored the Point Spread Function of the JPEG-2000 and JPEG transforms as a function of spatial position within a 2D image and was able to restore the compressed image to yield about 2dB improvement relative to the original image by deconvolution with the worst case Point Spread Function. I applied the Lucy-Richardson method implemented in the ESO-MIDAS data analysis package. Unsharp masking the original image yielded modest improvement in the visual quality of the compressed image.

JPEG blocking artifacts may be effectively suppressed by adding a small contribution from the Discrete Sine Transform<sup>13</sup>. Furthermore, the residual carries information about instrument noise, sky background and unresolved sources. Although Rice coding has been used effectively to encode residuals with distribution similar to Laplacian, an alternative approach is to report summary statistics that characterize the stochastic process of the residuals. Such statistics include the anisotropy or energy of the residuals (components of the structure tensor), minimum and maximum residual sky brightness or the parameters of an Autoregressive Moving Average model. Although it is often stated that noise is incompressible, identification of the noise process parameters conveys information to reconstruct the noise.

The *cfitsio* package<sup>14</sup> implements a tiled data compression scheme within the framework of the FITS standard. If data variables of differing statistical properties can be segmented into separate tiles, they can be approximated by methods optimized for such data.

## 4. QUALITY ASSESSMENT

I determined the visual quality resulting from application of the approximation techniques detailed in Table 3 and 4 to public domain images obtained via NASA's SkyView, the HST Heritage website together with full images from the SUBARU telescope observation archive. In particular, I focused on a Digitized Sky Survey image at 1024x1024 resolution centered on the Coma Cluster (Epoch J2000.0, Rectangular Equatorial coordinates, Gnomonic projection), a full-sky 10-band COBE/DIRBE image (Epoch J2000.0, Galactic coordinates, Zenith Equal Area projection) and an HST color image of the filamentary nebulosity surrounding the Cas A remnant.

Figure 2 below demonstrates the visual quality resulting from converting an 8-bit 1024x1024 Digitized Sky Survey image of the Coma Cluster from NASA's SkyView to a 4 bit-per-pixel image displayed by the FireView™ application at 160x120 resolution on a Palm IIIc emulator. The image is scrollable and zoomable

with size 187k, a compression factor of 5.6:1.

Figure 3 indicates the visual quality of an original full-sky map measured by the COBE/DIRBE mission. In particular, Figure 3(a) indicates the quality of a histogram-equalized Band 8 image raster. I decorrelated the spectrum via Principal Components Analysis of the image intensity in a 100x100x10 pixel window chosen arbitrarily at the Galactic Center. Although the spectral signature undoubtedly varies across the image and perhaps also with time, the color space resulting from this approximate decorrelation sufficed to obtain a high visual quality reconstruction from the single most energetic component spectral band. This provided a compression factor of 10:1 additional to that provided by compression in the spatial domain. Subsequent compression by a JPEG-2000 to 0.4bpp, 0.2bpp and 0.03 bpp resulted in the images displayed in Figure 3(b),(c) and (d). with an overall compression factor of 250:1. Fig. 3(e) shows the result of JPEG-2000 compression after spectral decorrelation at 1 bpp with additional deconvolution of the compressed image by the measured JPEG-2000 Point Spread Function and 3(f) shows the original image compressed to 0.03 bpp via the FRIT.

Figure 4 depicts three visual representations of an HST image of the Cas A remnant. Figure 4(a) shows the color image reduced by a factor of 2 in size and from 24-bit color to 16-bit gray scale of size 565 kB uncompressed (358K compressed via gzip) together with Figure 4(b) depicting a freehand artistic sketch of size 49 kB uncompressed (17 kB compressed via gzip) and Figure 4(c) illustrating a half-size freehand sketch of size 47kB uncompressed (1.8KB compressed via gzip). Even without line simplification, the sketches show a remarkable reduction in data volume while extracting useful topological information from the image. In the half-size sketch I have replaced the bright star images by points and reduced the polygonal outline of the filaments to a centerline. Active contours (e.g., snakes) show promise in extracting such information and further research work in this area would be of value to astronomers.

By providing simple representations of data that are not misleading to astronomers, it is possible to provide higher quality information with fewer bits of data. Although astronomers frequently refer to images as sky maps, cartographic techniques such as exaggeration, simplification, aggregation and displacement of features are rarely applied. I represent a crowded stellar field detected by a data reduction system such as ESO MIDAS by an oval or polygonal boundary with point sources such as bright stars and galaxies represented by points or symbols. Filaments are reminiscent of linear features and may be simplified like roads.

The extension of these concepts to higher dimensionality is interesting. I visualize data and models in 4D to 8D in terms of animating or color coding 3D models to represent 4D surfaces<sup>15</sup>, displaying down-projections of model subspaces or representing the spatial distribution of higher dimensional data in terms of the intensity of multiple speaker audio or other sensory stimuli. For very high dimensions, Donoho<sup>16</sup> points out that the concentration of measure phenomenon indicates that there is no harm in repeatedly searching data for interesting variables by minimizing the Residual Sum of Squares plus a penalty depending directly on the variance and model complexity (or number of free parameters) and logarithmically on the number of dimensions of the data. For raster data, it is straightforward to extend the FRIT, DPCM and JPEG-2000 and multiple dictionary LZW techniques to higher dimensions. For vector data, it may be especially helpful to encode the color or line-style coding of displayed contours or snakes for each dimension projected on the plane in the same manner as cartographic thematic layers. I prefer to model the residuals by evaluating the structure tensor or stochastic process coefficients to convey information useful to astronomers.

## 5.SUMMARY AND CONCLUSIONS

In summary, a system to view and analyze quick look astronomical data on hand held devices linked by wireless network to the Internet seems practical and cost-effective in comparison with systems based on laptop computers. Further research is required to extract and simplify astronomical information from large quantities of observational data and present it in a manner as compelling as Galileo's hand-drawn sketches.

## ACKNOWLEDGMENTS

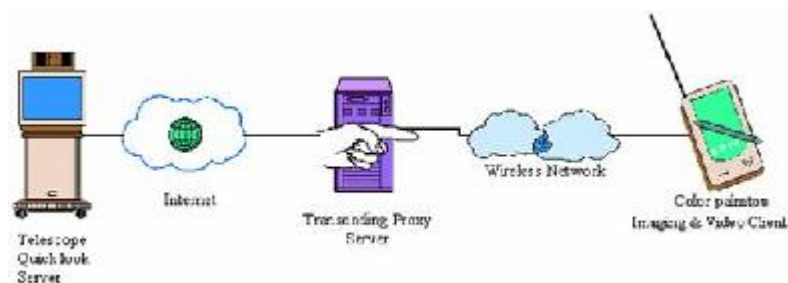
The COBE datasets were developed by the NASA Goddard Space Flight Center under the guidance of the COBE Science Working Group and were provided by the NSSDC. The Digitized Sky Survey was produced at the Space

Telescope Science Institute under U.S. Government grant NAG W-2166. The images of these surveys are based on photographic data obtained using the Oschin Schmidt Telescope on Palomar Mountain and the UK Schmidt Telescope. The plates were processed into the present compressed digital form with the permission of these institutions. The HST data are courtesy of NASA and the Space Telescope Science Institute.

This work was supported entirely by the internal funds of Dr. Immanuel Freedman, Inc.

## REFERENCES

1. M.Louys et al., "Astronomical Image Compression", *Astron. Astrophys. Suppl. Ser.* **136**, pp. 579-590, 1999
2. M. Hamabe et al., "New Image Quick-Look System for Subaru Telescope Data Archive", ASP Conf. Ser. **216**, *Astronomical Data Analysis Software and Systems IX*, eds. N. Manset, C. Veillet and D. Crabtree, p. 482, 1999
3. E.R. Tufte, *Envisioning Information*, Graphics Press, Cheshire CT 1990
4. M. Agrawala and C. Stolte, "Rendering Effective Route Maps: Improving Usability Through Generalization", *SIGGRAPH 2001, Computer Graphics Proc.*, ed. E. Flume, pp. 241-250, 2002
5. M. Jansen and M. van Kreveld, "Evaluating the Consistency of Cartographic Generalization", *Proc. 8<sup>th</sup> Intl. Sym. On Spatial Data Handling*, eds. T. K. Poiker and N. Chrisman, pp. 668-678, 1998
6. A. B. Watson, "Image data compression having minimum perceptual error", US Patent 5,426,512, 1995
7. D. Santa-Cruz, R. Grosbois and T. Ebrahimi, "JPEG 2000 performance evaluation and assessment", *Signal Processing: Image Communication* **17(1)**, pp. 113-120
8. M. N. Do and M. Vetterli, "The Finite Ridgelet Transform for Image Representation", Accepted *IEEE Transactions on Image Processing*, 2001
9. S. V. Vasilyev, "An Optimal Data Loss Compression Technique for Remote Surface Multiwavelength Mapping", ASP Conf. Ser. **145**, *Astronomical data Analysis and Software Systems VII*, eds. R. Albrecht, R. N. Hook and H. A. Bushouse, ASP, 1998
10. C. N. Sabbey, ASP Conf Ser. **172**, *Astronomical Data Analysis and Systems VIII*, eds. D. M. Mehringer, R. L. Plante and D. A. Roberts, ASP, San Francisco, 1999
11. C. Y. Xu and J. L. Prince, "Generalized gradient vector flow external forces for active contours", *Signal Processing* **171**, pp. 131-139, 1998
12. M. Bader and M. Barrault, "Improving Snakes for Linear Feature Displacement in Cartographic Generalization", *Geocomputation 2000*, 2000
13. P. M. Farrelle, *Recursive Block Coding for Image Data Compression*, Springer-Verlag, New York 1990
14. W. Pence, "CFITSIO v. 2.0, A Full-Featured Data Interface", ASP Conf. Ser. **172**, *Astronomical Data Analysis and Software Systems VIII*, p. 487, eds. D. Mehringer, R. Plante and D. A. Roberts, ASP, San Francisco, 1999
15. A. J. Hanson, K. I. Ishkov and J. H. Ma, "Meshview: Visualizing the Fourth Dimension", <http://citeseer.nj.nec.com/260456.html>
16. D. L. Donoho, "High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality", <http://www-stat.stanford.edu/seminars/seminars0001.html>



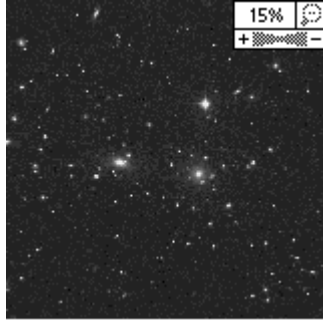
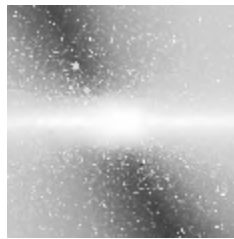


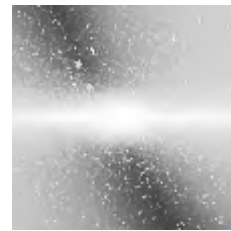
Fig 2. Scrollable and zoomable 1024x1024 image of the Coma Cluster from the Digitized Sky Survey from NASA's SkyView and displayed at 160x120x4 resolution by FireView™ on the Palm IIIc emulator



(a)



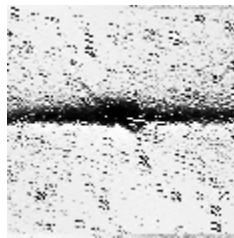
(b)



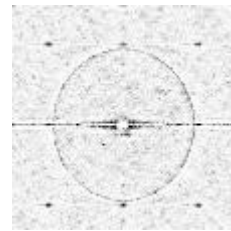
(c)



(d)



(e)



(f)

Fig 3(a). Histogram-equalized COBE/DIRBE Band 8 sky map. (b) spectrally decorrelated via PCA and compressed vi JPEG 2000 to 0.4 bpp, (c) 0.1 bpp, (d) 0.03bpp, (e) at 1 bpp with additional deconvolution of the compressed image by the measured JPEG-2000 Point Spread Function and (f) compressed at 0.03 bpp via the FRIT.



Fig 4(a) Cas A remnant observed by HST (358 kB).



Fig 4(b) Hand-drawn sketch of Cas A remnant (17 kB).



Fig 4( c) Simplified hand-drawn sketch of Cas A remnant at half the scale of Fig. 4(b) above.(1.8 kB).