DATA COMPRESSION FOR DATA ARCHIVAL, BROWSE OR QUICK-LOOK

Jeff Dozier
Universities Space Research Association
Goddard Space Flight Center
Greenbelt, MD 20771

James C. Tilton
Goddard Space Flight Center
Greenbelt, MD 20771

1. The Applications

1.1 Archival

Soon after space and Earth science data is collected, it is stored in one or more archival facilities for later retrieval and analysis. Since the purpose of the archival process is to keep an accurate and complete record of data, any data compression used in an archival system must be lossless, and protect against propagation of error in the storage media. In contrast, browse and quick-look require only the retrieval of a good approximation of the data, allowing consideration of lossy data compression. What is a good approximation depends, of course, on the data characteristics and the purposes for which the data is being browsed or previewed.

1.2 Browse

A browse capability for space and Earth science data is needed to enable scientists to check the appropriateness and quality of particular data sets before obtaining the full data set(s) for detailed analysis. Browse data produced for these purposes could be used to facilitate the retrieval of data from an archival facility. Appropriately derived browse data can also facilitate interdisciplinary surveys which search for evidence of unusual events in several data sets from one or more sensor. Such browse data can also be used to validate the quality of the data by facilitating quick checks for data anomalies.

1.3 Quick-look

Quick-look data is data obtained directly from the sensor for either previewing the data or for an application that requires very timely analysis of the space or Earth science data. This quick-look data could be either a small subsection of the full resolution data, or an approximate representation of a larger section of data, such as described for browse data. In the latter case, lossy data compression techniques tailored to retain the information significant to the particular application would be appropriate. Two main differences between data compression techniques appropriate browse and quick-look cases are the quick-look techniques (i) can be more specifically tailored, and (ii) must be limited in complexity by the relatively limited computational power available on space platforms.

2. Key Issues

2.1 Archival

Storage space: If lossless encoding is required, possible compression savings are limited to approximately 2:1 for most space and Earth science data. If this is the only justification for data compression, the use of data compression may not be justified since one could just buy twice as much of the storage media.

Data integrity: Any encoding of the data must be robust to errors in the storage media, and must retain the full scientific information content of the original data. For experimental data, this would generally mean that every bit of the original data must be retained.

Data access: Quick access is required to information about archived data, allowing interactive ordering of data from the archive. Appropriate browse data product(s) could serve to augment other descriptive data that is kept on-line for fast access, while the full data set is kept in off-line storage. Algorithms for decoding the compressed browse or full resolution data must be very fast. However, encoding speed is not critical, since there will be many decodes per encode.

Synergism: If decrease in storage space does not justify the use of data compression, a system employing data compression as an integral part that decreases storage space requirements, increases data integrity and improves data access would most certainly be justifiable.

2.2 Browse

Facilitate Access to Archived Data: Essential information for a wide variety of applications must be retained in the browse data for widest utility. A multitude of scientific data products may be generated from most space and Earth Science data sets. In addition, space and Earth Science data sets come in several different forms, including images, time series, 3 or 4-dimensional data, and housekeeping or ancillary data. For efficiency, browse data compression must be well integrated into the archival/data access facility. A well integrated browse facility would enable interactive ordering of archived data, and speed access over remote networks. In such a facility required information could be retained on-line for quick access.

Search for Unusual Events or Data Anomalies: Browse data produced by approaches that smooth the data too much, or bias towards expected or previously observed data signals, are not acceptable for these purposes.

Browse Data Quality: What quality is required? Can scientific analysis be performed on browse data? Can the production of browse data be made sufficiently "smart" to retain the information required for at least a preliminary scientific analysis of the data? The effects of the lossy compression used to produce the browse data must be analyzed for the effects on the results of the scientific analysis of the data (rather than just visual appearance).

Modes of Access: The user may want to be able to compare visually many browse images at one time, and then select one or more for more detailed analysis. Alternatively, the user may want to look at large portion of a data set in browse mode, and then focus done to a smaller subset for more detailed analysis.

2.3 Quick-Look

Computational Complexity: Quick-look can most easily be done as a rapid transmission at full resolution of a small subset of the data. When doing more than subsetting the data, the encoding algorithm must be limited in complexity by the relatively limited computational power available on space platforms. It is difficult to space qualify more powerful computer hardware.

Tailoring: Since quick-look data would be used for a specific purpose, the production techniques can be specifically tailored to the application.

2.4 Other

To facilitate wide participation in the development process, NASA data compression systems should follow accepted standards as closely as possible, such as JPEG (Joint Photographic Experts Group) or MPEG (Moving Picture Experts Group).

3. Data Compression Approaches

3.1 General Approaches

The data compression field is already highly developed. Given here, instead of a review of techniques, is a bibliography books on compression recommended provided by Robert M. Gray:

Lossless Data Compression (Noiseless Coding):

J. Storer, *Data Compression: Methods and Theory*, Computer Science Press, 1988.
T. J. Lynch, *Data Compression: Techniques and Applications*, Lifetime Learning, Wadsworth, 1985.

Transform and Predictive Coding:

N. S. Jayant, ed., *Waveform Quantization and Coding*, IEEE Press, 1976.
N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, 1984.
R. J. Clarke, *Transform Coding of Images*, Academic Press, 1985.
A. N. Netravali and B. G. Haskell, *Digital Pictures: Representation and Compression*, Plenum Press, 1988.

Vector Quantization:

H. Abut, ed., *Vector Quantization*, IEEE Press, 1990.
M. Rabbani and P. Jones, *Digital Image Compression*, SPIE Publications, 1991.
A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer, 1991.

3.2 Progressive Transmission

Progressive transmission techniques are a natural match to efficiently combining browse and data archival. Progressive transmission techniques can losslessly encode data, but the early stages of reconstruction naturally produce choices of data renditions that could be used as a browse version of the data. If none of the renditions is satisfactory as the browse version of the data, other means could be used to produce the browse version, and the difference between the browse data and original data could be losslessly compressed by progressive or other means. In either case only the information required to produce the browse rendition would be kept on-line, while the remainder of the information required to reproduce the original data would be retained in off-line storage.

3.3 Synergism with Analog to Digital (A-D) Conversion

Nearly all Space and Earth Science data collection involves A-D conversion. Since A-D conversion is in itself a gross form of lossy data compression, gains in information content per volume of data may be obtained by combining more sophisticated forms of lossy data compression with A-D conversion. The current approach using a uniform (or perhaps companded) quantizer for A-D conversion followed by lossless compression (if compression is employed) is suboptimal. An example of employing lossy compression techniques to optimize this process would be convert the analog signal into vector codes, such as done in vector quantization (a form of lossy compression). Vector quantization design techniques could then be employed to tailor the overall source code to characteristic of the data being encoded.

3.4 Other

If a large amount of on-board memory is available, a possible approach to data compression would be to just transmit the changes observed in the data from the same location from one orbit to the next.

Besides large amounts of on-board memory, this approach would require sufficient computational power to register the data collected in the current orbit with that from the previous orbit.

4. Open Questions

How predictable is a time series of images when the time interval is days, rather than seconds or split seconds? Can we losslessly compress a time series of, for example, MODIS data?

How can a browse system be designed intelligently so various types of remote sensing data (SAR data, multi-spectral data, or spectrometer data), time series data (with small time intervals), or housekeeping/ancillary data are handled appropriately?

5. Recommendations

There is a critical need to promote interaction between data compression scientists and space and Earth scientists to more effectively explore the utility of data compression techniques for space and Earth science data. A first step that can be done immediately (without specific new funding) is for NASA to provide test data sets and examples of analysis scenarios to data compression scientists. This data and scenario information could be kept at an "anonymous ftp" site, and/or made available on an optical disk. At a minimum, this will enable researchers to determine if their existing techniques are, or are not, appropriate for space and Earth science data. A more structured (i. e., funded) program would be required to insure feedback and more intensive refinement of approaches to suit the data and analysis scenarios. Possibly this effort could tap into the Version 0 EOSDIS activity. An important task to be accomplished by a more structured activity would be to statistically characterize the various classes of space and Earth data.

Certain technical approaches stand out as being particularly promising. The application of data compression to browse and data archival is one. Development of this type of system for various data types should be promoted. Also to be encouraged is the production of "smart" browse data for various different data types and applications. This "smart" browse data would retain most of the essential information for a rough, but still informative, scientific analysis of the data. This research would provide feedback concerning the best types of browse data to provide as an integral part of a data archival access system.

Another area of research that should be encouraged is the combination of lossy compression techniques with analog to digital (A-D) conversion.

We recommend that NASA should make the pursuit of research in these and other promising areas related to the compression of space and Earth science data an area of emphasis in one or more future solicitations (e.g., NASA Research Announcement) under the Applied Information Systems Research Program and/or other appropriate NASA program.

The organizers of the Data Compression Conference, of which this workshop is a part, have already announced that the next Data Compression Conference (DCC'92) will be held on March 24-27, 1992 in Snowbird, Utah. We recommend that participants in DCC'92 be encouraged to test their methods on a standard set of images provided by NASA. This standard set of images might include Landsat Thematic Mapper images, AVIRIS images, SAR images, space time series data. Perhaps some "bad" data should also be included. A special session at DCC'92 could be devoted to discussing and contrasting these results.

Participants

The participants in this discussion group were Karen Anderson, Joe Bredekamp, Mayun Chang, Pamela C. Cosman, Linda Jo Dolny, Jeff Dozier, Benjamin Epstein, Wai-Chi Fang, Robert G. Finch, Richard Frost, Daniel Glover, Robert Gray, Paul G. Howard, Peter Kenny, Manohar Mareboyana, Kristo Miettinen, Jeff Niehaus, Karen Oehler, Dale Rickman, Eve Riskin, S. Srikanth, Vincent Salomonson, Rick Schumeyer, James C. Tilton, Jeff Vitter, and Ray Walker. See the appendix for addresses.