A Hierarchical Storage Management (HSM) Scheme for Cost-Effective On-Line Archival Using Lossy Compression

David E. Avrin, Katherine P. Andriole, Lloyd Yin, Robert G. Gould, and Ronald L. Arenson

A hierarchical storage management (HSM) scheme for cost-effective on-line archival of image data using lossy compression is described. This HSM scheme also provides an off-site tape backup mechanism and disaster recovery. The full-resolution image data are viewed originally for primary diagnosis, then losslessly compressed and sent off site to a tape backup archive. In addition, the original data are wavelet lossy compressed (at approximately 25:1 for computed radiography, 10:1 for computed tomography, and 5:1 for magnetic resonance) and stored on a large RAID device for maximum cost-effective, on-line storage and immediate retrieval of images for review and comparison. This HSM scheme provides a solution to 4 problems in image archiving, namely cost-effective on-line storage, disaster recovery of data, off-site tape backup for the legal record, and maximum intermediate storage and retrieval through the use of on-site lossy compression.

KEY WORDS: PACS, computer, archiving, radiology, compression, wavelet, hierarchical storage management.

IT IS ESSENTIAL to maximize on-site storage at minimum cost for digital archival of images to be clinically and economically beneficial. In a 1998 report, Pratt et al[1] commented that the high cost of PACS (picture archiving and communication system) was largely because of the need for multiple expensive jukebox storage devices. The application service provider (ASP) approach to archiving, in which storage of data and maintenance of hardware devices is performed by a provider, is currently too costly, particularly for large-volume sites. The hierarchical storage management (HSM) scheme presented here maximizes on-line storage for immediate image retrieval and overall cost effectiveness using wavelet compression. This solution also provides for disaster recovery with the legal record maintained losslessly compressed and stored on an inexpensive digital linear tape (DLT) archive.

Wavelet encoding has shown great promise for medical imaging.[2] Visually lossless but numerically lossy compression appears to be possible at compression ratios of between 5:1 and 30:1 or more, depending on the imaging modality.[3,4] In the method presented here, the original DICOM (Digital Imaging and Communications in Medicine Standard) image data are used for primary interpretation and are maintained locally on a large RAID (redundant array of independent disks) device for several weeks, whereas lossy compressed images are kept online or nearline as reference and historical images. The application of non–bit-preserving compression at the stated ratios can significantly grow the on-line storage capacity while still maintaining adequate quality images for review or comparison.

A losslessly encoded copy of the data is transferred to an offsite tape archive thus making use of inexpensive media for the permanent lossless record to satisfy any legal requirements. It also serves as a backup, and with an associated database, a resource for disaster recovery. It is interesting to note that in general, backup copies are not expected or required of a conventional film archive. In the digital world, however, backup procedures are considered to be part of standard data processing practice.

MATERIALS AND METHODS

Figure 1 diagrams the HSM scheme, which currently provides the back-end archival services for a conventional Agfa Version 3.5 PACS. Image and relevant data are acquired into the PACS through a network gateway, and demographic data are verified with that in the Radiology Information System (RIS). The original DICOM image data are routed to a 236 gigabyte (GB) RAID for current, recent, and active patients, and to 1 or more workstations for primary diagnosis. The original DICOM data are held in the on-site RAID for approximately 2 weeks and autodeleted using a first-in, first-out scheme. Data also are transmitted to the University of California at San Francisco (UCSF) parallel PACS that provides an environment for implementing and evaluating the HSM.

The original DICOM data undergo a wavelet encoding process[5] at a targeted lossless compression ratio of 2.5 or 3:1 within 4 hours of being acquired. The algorithm used and cited in the reference[5] allows the specification of a target compression ratio, which optimizes the encoding process for that ratio. The wavelet encoding and decoding is accomplished using the MrSID (Multiresolution Seamless Image Database) software

From the Department of Radiology, University of California San Francisco, San Francisco, CA.

Address reprint requests to David Avrin, MD, PhD, Department of Radiology, University of California San Francisco, 505 Parnassus Ave, Room 392, San Francisco, CA 94143-0628.

0897-1889/01/1401-0004$35.00/0

doi:10.1053/jdim.2001.21441

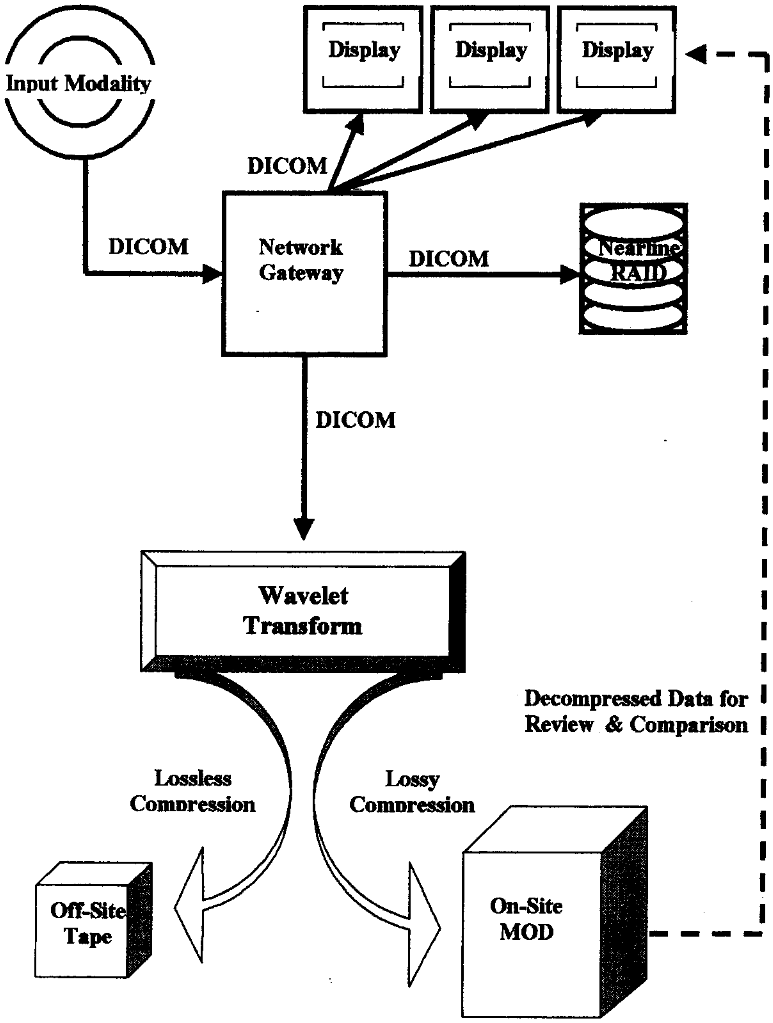

Fig 1. HSM Scheme. Image data are viewed at original content for primary diagnosis, losslessly compressed for the off-site legal record on tape, and wavelet lossy compressed for the on-site near-line storage on MOD for review and historic comparison.

component supplied by LizardTech (Seattle, WA). This data set is passed continuously via a conventional 10 megabit per second Ethernet in chronological order to a simulated offsite DLT jukebox and controlling serving with an independent database. Once the on-site system receives confirmation that the off-site data reception and storage commitment is error free, then lossy wavelet encoding at higher ratios is performed. Target compression ratios of 5 to 10:1 are used for the cross-sectional modalities and 20 to 30:1 for digital projection radiographs. The wavelet compressed data are committed to the 7 terabyte (TB) MOD (magneto-optical disk) jukebox for long-term on-line or near-line storage. Retrieval and prefetching of historic studies occurs from the compressed MOD archive.

The connection between the PACS and tape storage system is 10-megabit Ethernet. Individual cases can be retrieved from the off-site storage and decoded to their original digital content on demand if such an original representation is required. This occurrence is expected to be very rare. In the event of a technical or natural disaster, entire time ranges of image data can be repopulated from the off-site HSM component to the on-site PACS.

RESULTS AND DISCUSSION

Image Quality by Modality and Compression Ratio

Image quality was assessed subjectively and objectively. Processed images were compared with the originals by absolute pixel difference images for various modalities. RMS (root-mean-square) error statistics were calculated on the images as a function of compression ratio and modality.

Although mathematically difficult to prove, it appears that wavelet encoding, at least by this algorithm, is indeed lossless or bit-preserving for medical images of computed tomography (CT), magnetic resonance (MR), and computed radiography (CR) when compressed by a ratio of 2.5 or less. This is based on experimental tests resulting

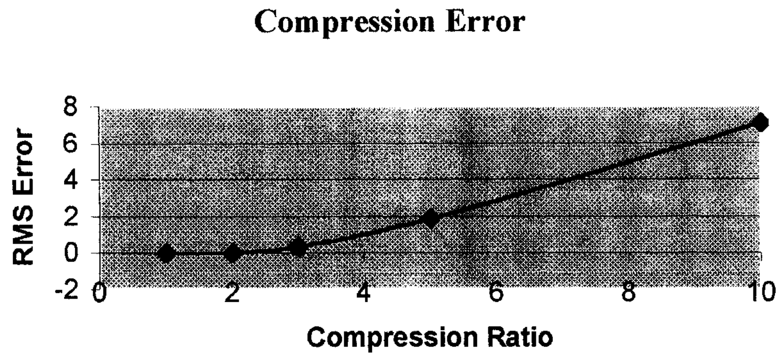

Fig 2. A graph of the root mean square compression error for a typical abdominal CT image as a function of compression ratio. Experimental tests show that at ratios of 2.5 and below, the compression is lossless (RMS Error = 0).

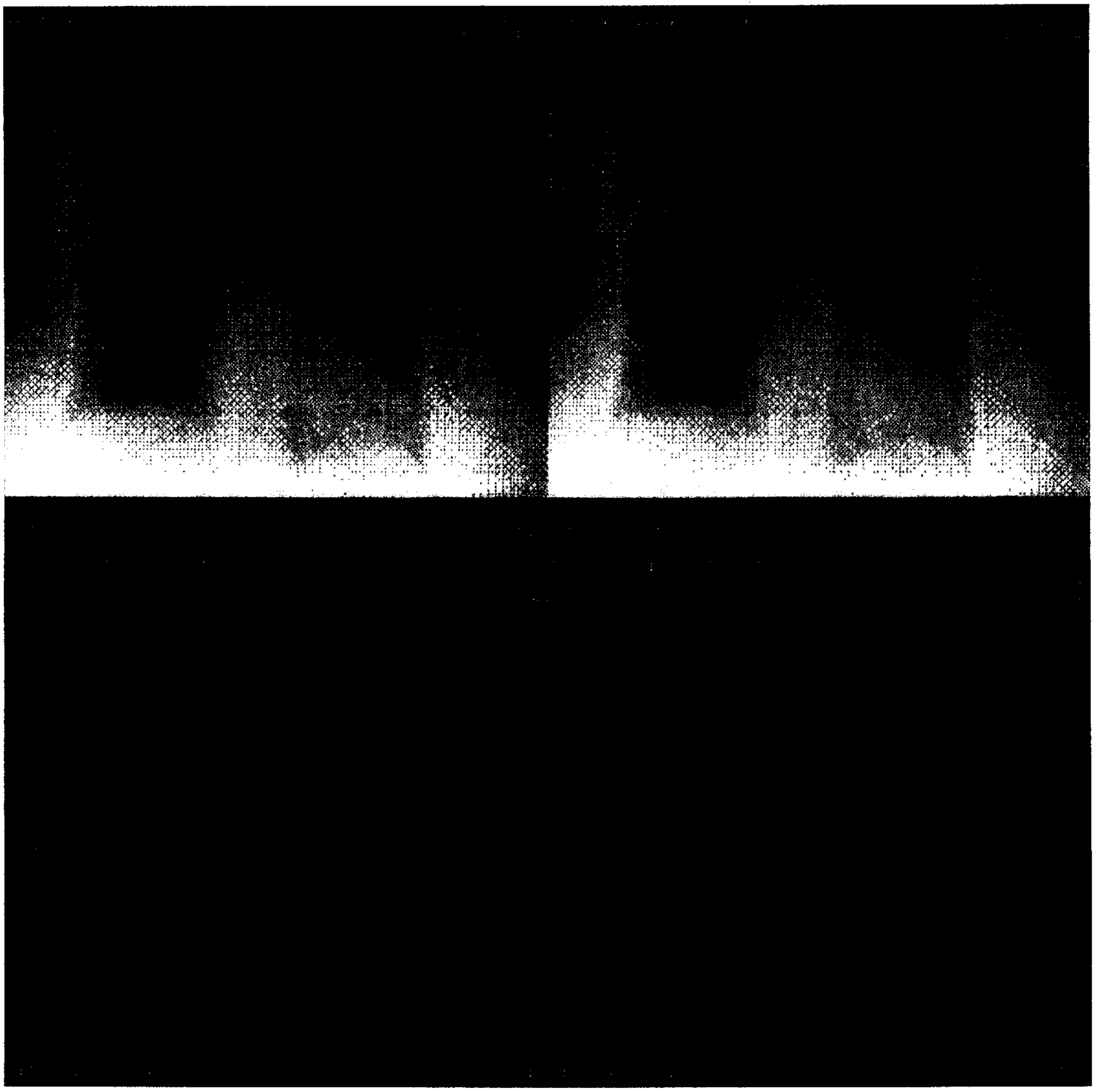

Fig 3. Original CR image (top, left), CR image wavelet compressed 30:1 (top, right), absolute difference image (bottom, left), and signed difference image (bottom, right).

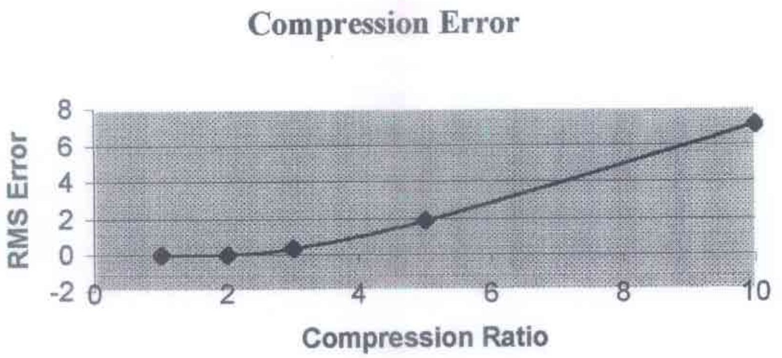

Fig 2. A graph of the root mean square compression error for a typical abdominal CT image as a function of compression ratio. Experimental tests show that at ratios of 2.5 and below, the compression is lossless (RMS Error = 0).

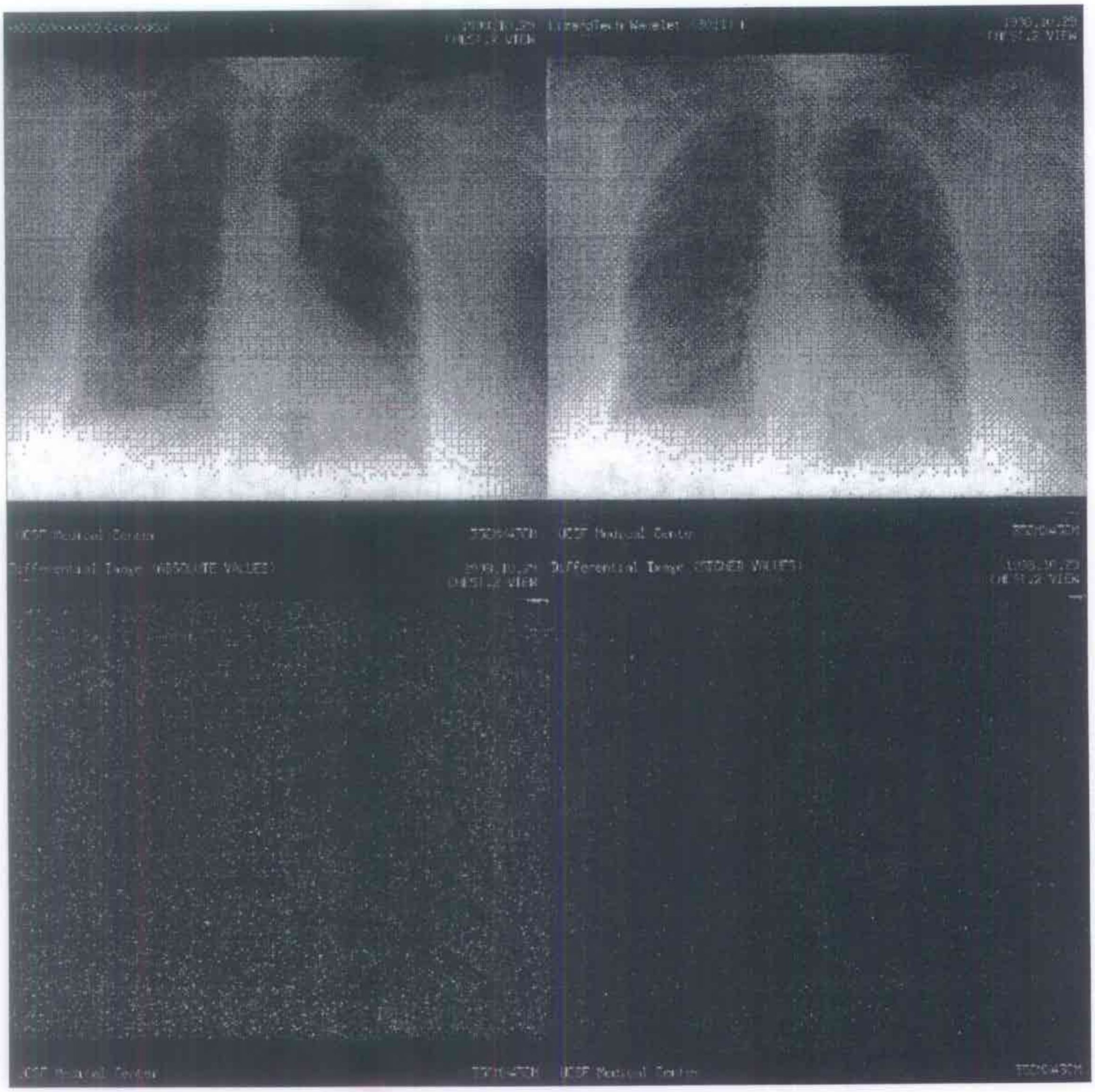

Fig 3. Original CR image (top, left), CR image wavelet compressed 30:1 (top, right), absolute difference image (bottom, left), and signed difference image (bottom, right).

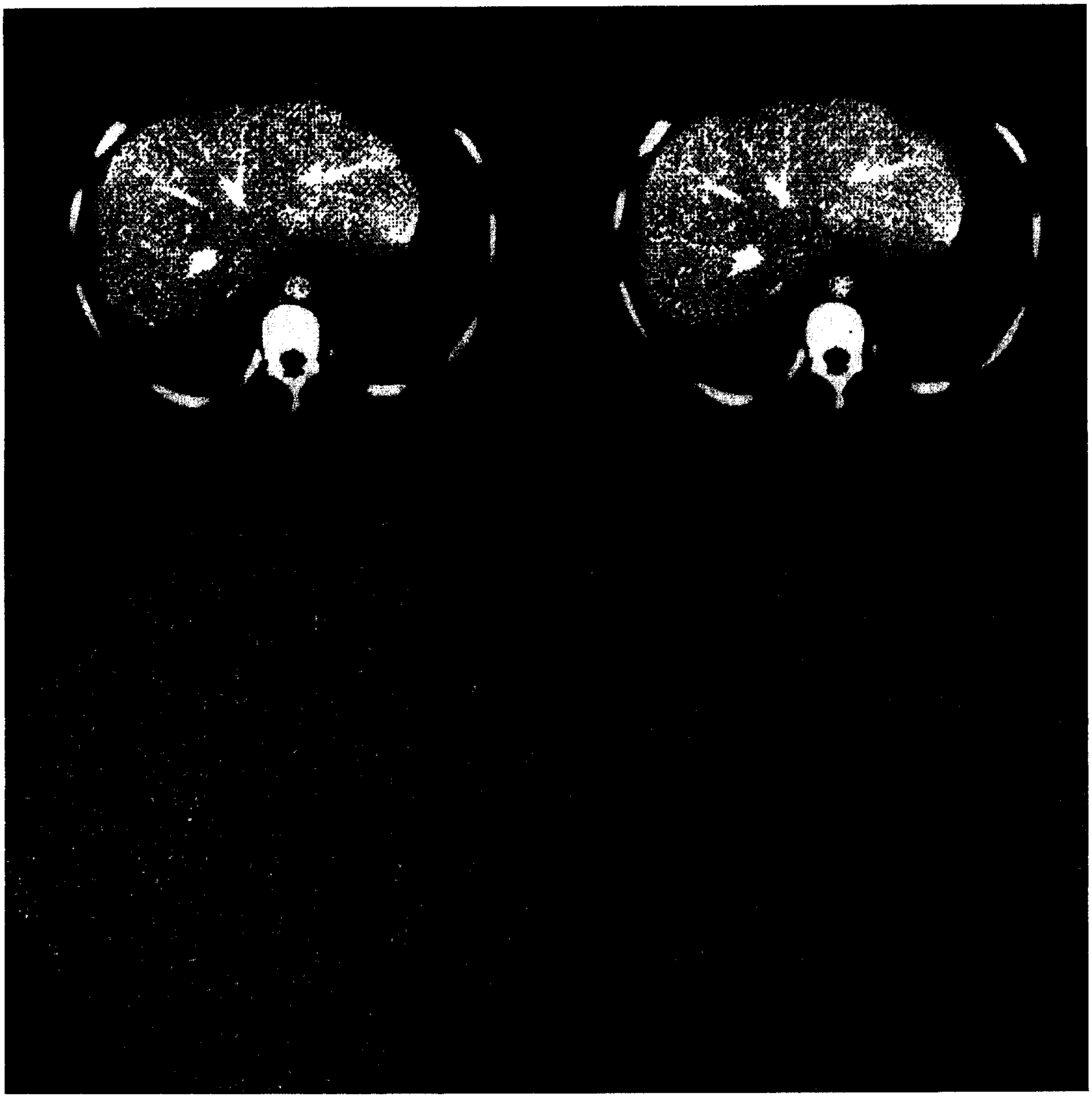

Fig 4. Original CT image (top, left), CT image wavelet compressed 10:1 (top, right), absolute difference image (bottom, left), and signed difference image (bottom, right).

in a RMS error equal to zero at those levels. An example pixel error graph of RMS error versus compression ratio is shown in Fig 2 for representative CT images.

Example compressed images for CR at 30:1, CT at 10:1, and MR at 5:1 are shown in Fig 3-5, respectively, in comparison with the original. Also shown is the absolute difference image calculated as the absolute value of the pixel difference between the reconstruction of the encoded image and the original image, and the signed difference image. Note that very little important anatomic structure is discernable in the difference images, which represent what is lost in compression of the data.

A graph of the pixel deviations for CT images as related to pixel frequency is shown in Fig 6 for compression ratios of 2:1, 3:1, 5:1, and 10:1. Note that if there is a deviation in pixel value between the original and the compressed image data, it is a very small deviation and realistically on the order

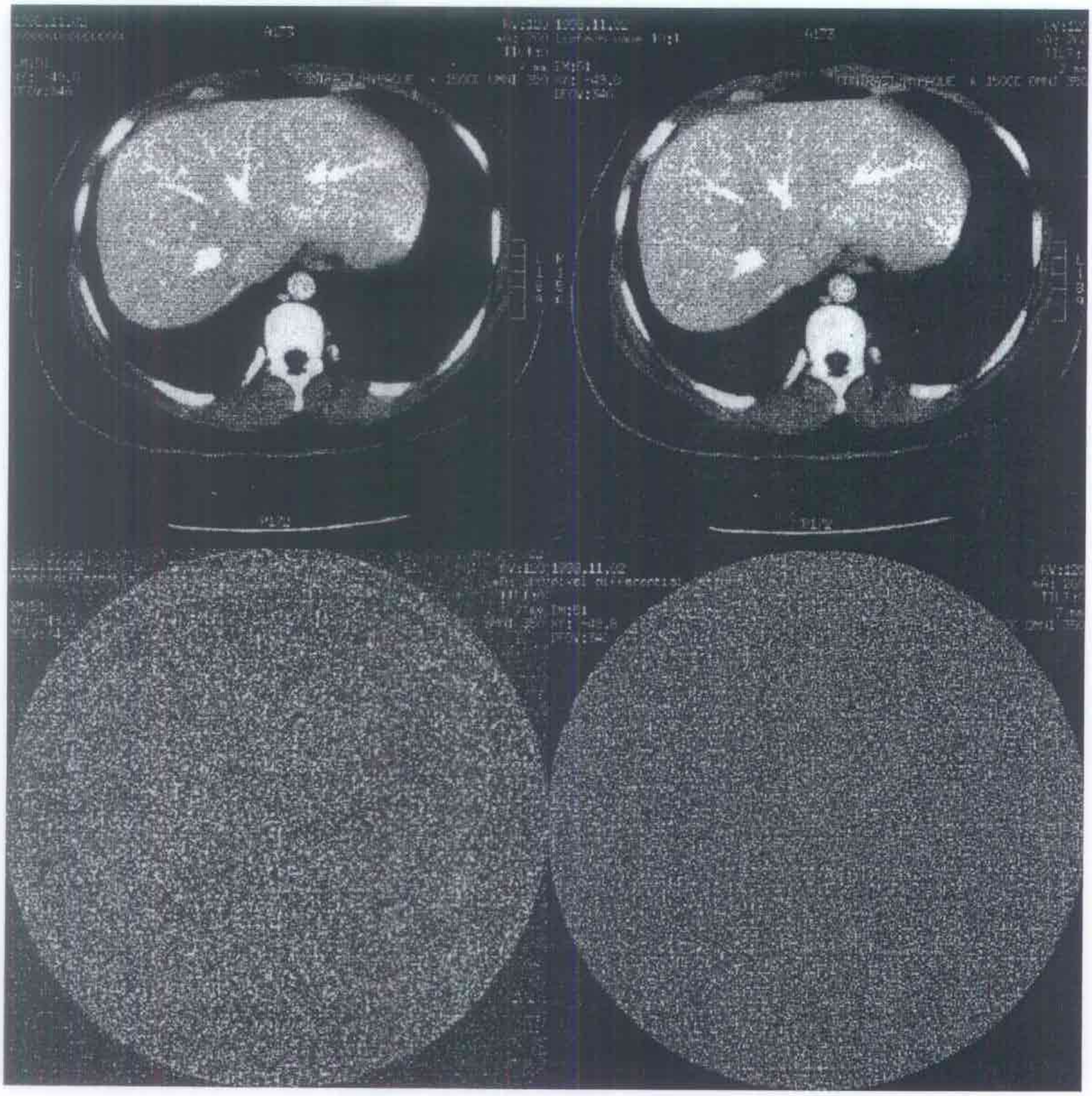

Fig 4. Original CT image (top, left), CT image wavelet compressed 10:1 (top, right), absolute difference image (bottom, left), and signed difference image (bottom, right).

in a RMS error equal to zero at those levels. An example pixel error graph of RMS error versus compression ratio is shown in Fig 2 for representative CT images.

Example compressed images for CR at 30:1, CT at 10:1, and MR at 5:1 are shown in Fig 3-5, respectively, in comparison with the original. Also shown is the absolute difference image calculated as the absolute value of the pixel difference between the reconstruction of the encoded image and the original image, and the signed difference image. Note that very little important anatomic structure is discernable in the difference images, which represent what is lost in compression of the data.

A graph of the pixel deviations for CT images as related to pixel frequency is shown in Fig 6 for compression ratios of 2:1, 3:1, 5:1, and 10:1. Note that if there is a deviation in pixel value between the original and the compressed image data, it is a very small deviation and realistically on the order

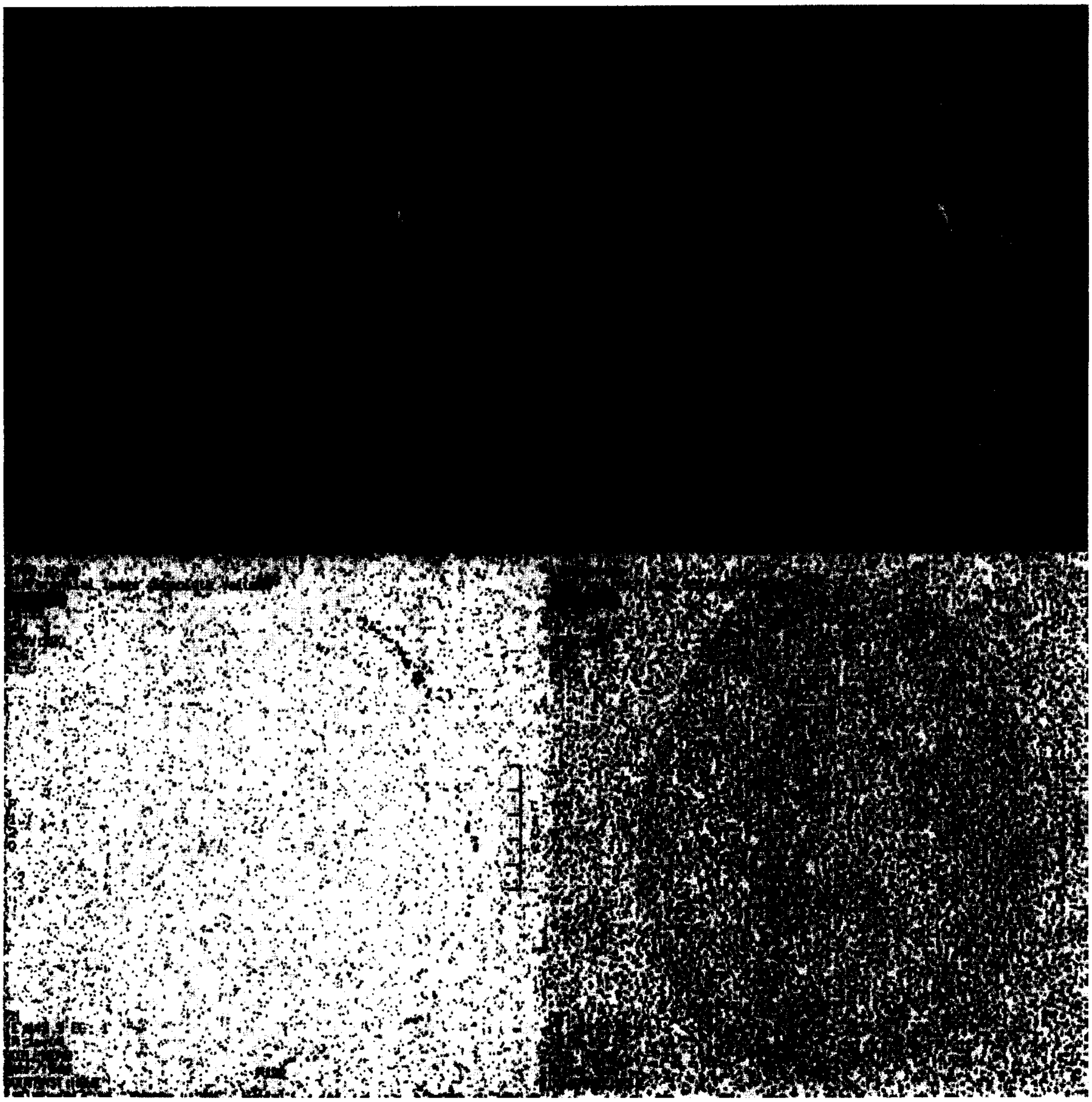

Fig 5. Original MR image (top, left), MR image wavelet compressed 5:1 (top, right), absolute difference image (bottom, left), and signed difference image (bottom, right).

of the noise in the image. Thus, the compressed images are not equivalent mathematically to the original data but visually and diagnostically very much like the original, particularly when used as reference images.

Archive Capacity Enhancement

With a typical case mix and size of 20% MR examinations at 25 MBytes per examination (200 images), 30% CT examinations at 40 MBytes per examination (80 images), and 50% CR examinations at 30 MBytes per examination (3 images), utilizing the compression ratios above, overall archive compression of approximately 10:1 can be achieved as compared with raw image files. (Compared with lossless compression at 2 or 2.5:1, lossy compression grows the archive by a factor of 4 or 5.) Thus, the 7 TB MOD jukebox, which holds 2 years of raw data, effectively becomes 70 TB and can store 7,000 days of studies at the current data acquisition rate of 10 GB per day. This accounts for approximately 20 years worth of image data, effectively maintaining the legal

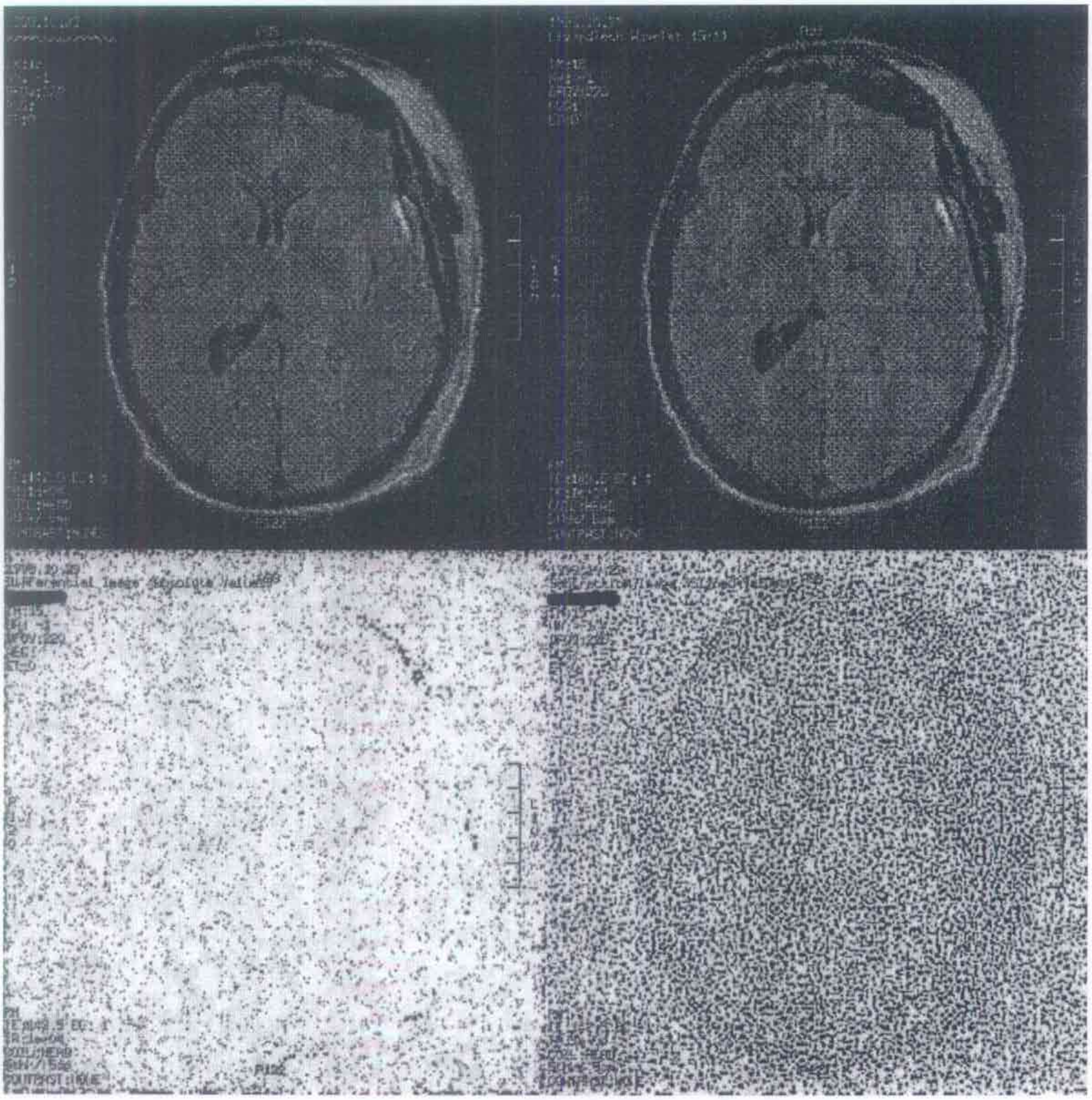

Fig 5. Original MR image (top, left), MR image wavelet compressed 5:1 (top, right), absolute difference image (bottom, left), and signed difference image (bottom, right).

of the noise in the image. Thus, the compressed images are not equivalent mathematically to the original data but visually and diagnostically very much like the original, particularly when used as reference images.

Archive Capacity Enhancement

With a typical case mix and size of 20% MR examinations at 25 MBytes per examination (200 images), 30% CT examinations at 40 MBytes per examination (80 images), and 50% CR examinations at 30 MBytes per examination (3 images), utilizing the compression ratios above, overall archive compression of approximately 10:1 can be achieved as compared with raw image files. (Compared with lossless compression at 2 or 2.5:1, lossy compression grows the archive by a factor of 4 or 5.) Thus, the 7 TB MOD jukebox, which holds 2 years of raw data, effectively becomes 70 TB and can store 7,000 days of studies at the current data acquisition rate of 10 GB per day. This accounts for approximately 20 years worth of image data, effectively maintaining the legal

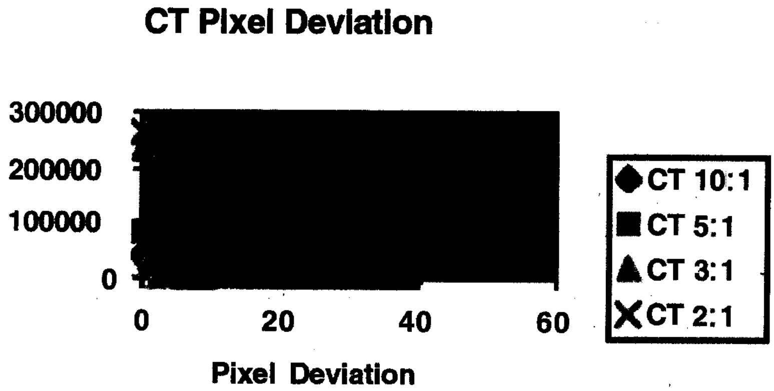

Fig 6. A graph of the pixel deviations for CT images as related to pixel frequency for compression ratios of 2:1, 3:1, 5:1, and 10:1. At 2:1, there is no error. At 3:1 a few pixels (of 256,000) differ by 2 Hounsfield units, and about 20,000 (4%) differ by only 1 Hounsfield unit.

record worth of original plus 2 relevant prior examinations all on line.

Alternatively, this increased space could be used to tailor the archive to given sites, needs, and resources. For example, a site high in brain tumor cases could use the space more effectively to keep more priors on line. Or a site following pediatric patients could maintain cases for the course of childhood.

Wavelet Encoding and Off-Site Transfer Rates

Wavelet encoding time with the components used is approximately linearly proportional to the file size (number of pixels) of the image. On the Sun Ultra 2 platform, with the version 2 of the MrSID encoder, single image encoding times are 3 seconds for a 0.5 MB CT slice, 1 second for a .125 MB slice MR, and 24 seconds for a 10 MB single image CR. This yields a worst case scenario of 8 secs/MB or 80,000 seconds for 10 GB of raw image data per day.

Continuous data transfer to the offsite archive is required at a 10 MB Ethernet transfer ratio to keep up with the flow of incoming data.

CONCLUSION

An HSM scheme utilizing short-term archival of uncompressed DICOM data (for primary diagnosis) in an on-site RAID, coupled with a very deep long-term archive of diagnostic quality wavelet compressed data in an on-site optical disk jukebox, cost effectively maximizes on-line storage for immediate image retrieval. Diagnostically lossy compressed data (at ratios of approximately 25:1 for CR, 10:1 for CT, and 5:1 for MR) grows the on-site jukebox by 10 times raw, depending on the mix of cases, making 20 or more years available on-line, effectively maintaining the legal record worth of original plus 2 relevant prior examinations all on line.

The off-site tape archival system provides data redundancy for backup and disaster recovery on a very low cost medium. It also maintains the legal record uncompressed, which can be retrieved as necessary from the off-site tape storage, where it is safe from simultaneous natural or technical disaster.

This HSM scheme provides a solution to 4 problems in image archival. Namely, cost-effective, on-line storage; disaster recovery of data; off-site tape backup for the legal record; and maximum intermediately storage and retrieval through the use of on-site lossy compression. It also allows a site to tailor an archive to their specific needs and resources, and may be more cost effective than the currently popular ASP model for archiving, particularly for high-volume sites.

REFERENCES

1. Pratt HM, Langlotz CP, Feingold ER, et al: Incremental cost of department-wide implementation of a PACS and CR. Radiology 206:245-252, 1998

2. Erickson BJ, Manduca A, Palisson P, et al: Wavelet compression of medical images. Radiology 206:599-607, 1998

3. Avrin DE, Andriole KP, Yin L, et al: A Hierarchical Storage Management (HSM) Solution for Cost-Effective On-Line Archival and Disaster Recovery Using On-Site Lossy Compression and Off-Site Tape Backup. 85th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, Nov 28-Dec 3, 1999

4. Savcenko V, Erickson BJ, Palisson P, et al: Detection of subtle abnormalities on chest radiographs after irreversible compression. Radiology 206:609-616, 1998

5. Mr Sid: Multi-resolution seamless image database. Seattle, WA, Lizardtech Corp (product brochure)

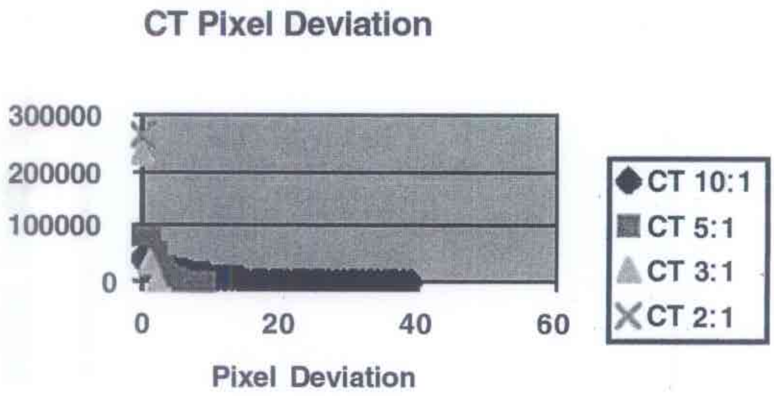

Fig 6. A graph of the pixel deviations for CT images as related to pixel frequency for compression ratios of 2:1, 3:1, 5:1, and 10:1. At 2:1, there is no error. At 3:1 a few pixels (of 256,000) differ by 2 Hounsfield units, and about 20,000 (4%) differ by only 1 Hounsfield unit.

record worth of original plus 2 relevant prior examinations all on line.

Alternatively, this increased space could be used to tailor the archive to given sites, needs, and resources. For example, a site high in brain tumor cases could use the space more effectively to keep more priors on line. Or a site following pediatric patients could maintain cases for the course of childhood.

Wavelet Encoding and Off-Site Transfer Rates

Wavelet encoding time with the components used is approximately linearly proportional to the file size (number of pixels) of the image. On the Sun Ultra 2 platform, with the version 2 of the MrSID encoder, single image encoding times are 3 seconds for a 0.5 MB CT slice, 1 second for a .125 MB slice MR, and 24 seconds for a 10 MB single image CR. This yields a worst case scenario of 8 secs/MB or 80,000 seconds for 10 GB of raw image data per day.

Continuous data transfer to the offsite archive is required at a 10 MB Ethernet transfer ratio to keep up with the flow of incoming data.

CONCLUSION

An HSM scheme utilizing short-term archival of uncompressed DICOM data (for primary diagnosis) in an on-site RAID, coupled with a very deep long-term archive of diagnostic quality wavelet compressed data in an on-site optical disk jukebox, cost effectively maximizes on-line storage for immediate image retrieval. Diagnostically lossy compressed data (at ratios of approximately 25:1 for CR, 10:1 for CT, and 5:1 for MR) grows the on-site jukebox by 10 times raw, depending on the mix of cases, making 20 or more years available on-line, effectively maintaining the legal record worth of original plus 2 relevant prior examinations all on line.

The off-site tape archival system provides data redundancy for backup and disaster recovery on a very low cost medium. It also maintains the legal record uncompressed, which can be retrieved as necessary from the off-site tape storage, where it is safe from simultaneous natural or technical disaster.

This HSM scheme provides a solution to 4 problems in image archival. Namely, cost-effective, on-line storage; disaster recovery of data; off-site tape backup for the legal record; and maximum intermediately storage and retrieval through the use of on-site lossy compression. It also allows a site to tailor an archive to their specific needs and resources, and may be more cost effective than the currently popular ASP model for archiving, particularly for high-volume sites.

REFERENCES

1. Pratt HM, Langlotz CP, Feingold ER, et al: Incremental cost of department-wide implementation of a PACS and CR. Radiology 206:245-252, 1998

2. Erickson BJ, Manduca A, Palisson P, et al: Wavelet compression of medical images. Radiology 206:599-607, 1998

3. Avrin DE, Andriole KP, Yin L, et al: A Hierarchical Storage Management (HSM) Solution for Cost-Effective On-Line Archival and Disaster Recovery Using On-Site Lossy Compression and Off-Site Tape Backup. 85th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, Nov 28-Dec 3, 1999

4. Savcenko V, Erickson BJ, Palisson P, et al: Detection of subtle abnormalities on chest radiographs after irreversible compression. Radiology 206:609-616, 1998

5. Mr Sid: Multi-resolution seamless image database. Seattle, WA, Lizardtech Corp (product brochure)