Compression Guidelines for Diagnostic Telepathology

David J. Foran, *Member, IEEE*, Peter P. Meer, *Senior Member, IEEE*, Thomas Papathomas, *Senior Member, IEEE*, and Ivan Marsic, *Member, IEEE*

***Abstract*— As the healthcare community has begun to rely increasingly upon digital technologies for acquisition, storage, and transmission of pictorial data, image compression has become an indispensable tool. We have investigated the feasibility of lossy compression in a well-defined task domain, the clinical assessment of digitized images of chromatic microscopic pathology specimens. The effect of compression was measured under two distinct perceptual criteria, just noticeable difference (j.n.d.) and largest tolerable distortion (l.t.d.), differing in the involvement required from subjects, who were experts in pathology. For standard JPEG compressed images it was found that when the experiment is performed under the l.t.d. criterion, a significantly larger compression ratio is reported as satisfactory. It is concluded that lossy compression holds promise for diagnostic telepathology.**

I. Introduction

THE DRAMATIC growth of the Internet in the last few years will certainly lead to significant changes in how healthcare is delivered [8]. As computer technology becomes ubiquitous, leading healthcare providers have begun to explore the use of digital images for rendering diagnostic decisions. Applying modern image understanding techniques to the analysis of digital medical images can sometimes provide the physician with important diagnostic information, otherwise unavailable without surgical intervention [5]. Picture Archiving and Communication Systems (PACS) are being utilized increasingly to manage clinical image data and to provide timely access to referring physicians [7]. Digital images, however, require large storage spaces which can easily exceed 20 Tbytes per year for a 1500-bed university hospital [16].

To manage such large amounts of data, the images are often stored in compressed format. Similarly, using compression can reduce the bandwidth requirements for transmitting images over networks. Advances in image processing technology allow one to compress or decompress large (1024×1024) images practically in real-time using specialized hardware. The goal of compression algorithms is to reduce the size of images while preserving their informational content.

There are two distinct classes of image compression methods: lossless and lossy. Lossless compression methods are fully reversible, but are typically confined to compression ratios on the order of 3:1. Lossless compression is not a practical solution when storage or transmission of large amounts of data is needed. Moreover, in digital imagery, quantization of the analog input information already introduces an irreversible processing step. In lossy image compression, the achievable compression rate is bounded by the tolerable degradation in image quality. See [16] for a review.

Reliable quantitative assessment of the quality of a compressed image is already an open problem in an engineering context, e.g., [6], [12]. In the medical domain, the use of image compression often raises legal issues [16]. Three different (although unrelated) measures can be defined to characterize the distortion of a compressed medical image relative to the original [3]:

- *Signal-to-noise ratio*, which is an objective measure of the distortion. Unfortunately describing the quality of an image with a single number is rarely sufficient.
- *Subjective rating*, which characterizes the subjective quality of the compressed image assessed on an ordinal scale.
- *Diagnostic accuracy*, which measures the perceived degradation through the usefulness of the image in clinical decision-making. Note the difference from the previous criterion.

Numerous studies have investigated aspects of image quality assessment for compressed medical images. See [3], [16], and references therein. However, the majority of these studies focus on gray-level images generated for radiology, computerized tomography and magnetic resonance applications.

In this paper we address the problem of image quality in the rapidly emerging field of telepathology. With the increased specialization of the pathologists, situations often arise in which consultation with an expert located at a distance is required [2]. Electronic transmission of the image of a specimen significantly reduces the time for getting an expert opinion, relative to the traditional approach of sending the specimen by mail. Current research in telepathology deals with the issue of diagnostic accuracy achieved with video microscopy versus performance with light microscopy. Studies were performed for a wide range of operating conditions, from static images to those driven by robotic video microscope systems. See [2], [9], and [10].

Manuscript received November 5, 1996; revised January 7, 1997. This work was supported by the New Jersey Commission on Science and Technology through Grant CAIP 96-06.

D. Foran is with the Center for Biomedical Imaging, UMDNJ-Robert Wood Johnson Medical School, Piscataway, NJ 08854 USA (e-mail: djf@pleiad.umdnj.edu).

P. P. Meer is with the Department of Electrical and Computer Engineering and the Center for Computer Aids for Industrial Productivity, Rutgers University, Piscataway, NJ 08855 USA.

T. Papathomas is with the Laboratory of Vision Research, and Department of Biomedical Engineering, Rutgers University, Piscataway, NJ 08855 USA.

I. Marsic is with the Center for Computer Aids for Industrial Productivity, Rutgers University, Piscataway, NJ 08855 USA.

Publisher Item Identifier S 1089-7771(97)03559-0.

Color is being utilized increasingly in telemedicine applications [15]. In diagnostic pathology applications for example tissue and molecular differences are often accentuated using specialized chromatic dyes and fluorescent agents while accurate digital representation of a single chromatic pathological specimen taken at high power can require as much as 7 GB, all of the relevant information for a radiology study can be captured in 25 MB [2]. There is a large redundancy in pathology images because the underlying staining process produces a relative restricted palette of colors.

The goal of our study was to establish a set of guidelines on achievable compression ratios in practical telepathology systems with emphasis on anatomic pathology and hematopathology. Two subjective measures were employed:

- assessment of the just noticeable difference (j.n.d.) in image quality degradation;
- the largest tolerable distortion (l.t.d.) for maintained *diagnostic accuracy*.

Both measures were quantized by the corresponding compression ratio thresholds. In an engineering context the two measures are related to the concepts of *image fidelity* and *image quality* between which a complex task-dependent relationship exists [13].

II. Experimental Design

A. Stimuli

The stimuli were generated from the slide archives of the Pathology Department of Robert Wood Johnson University Hospital, New Brunswick, NJ, and were a subset of an ensemble used in a previous experiment investigating the feasibility of a distributed, telepathology system [4].

In that experiment two senior residents randomly selected slides corresponding to 45 routine surgical pathology cases from the department's slide archives in New Brunswick, and ten cases from the archives in Hamilton, NJ. Five to 15 optical fields at various magnifications were selected, and the slides were digitized with a three-chip high-resolution SONY 970MD color camera interfaced with a Coreco Occulus acquisition board. The board generated both 24- and 8-bit color images using proprietary algorithms for color reduction.

Six staff pathologists reviewed the digitized cases as they were transmitted across computer networks and/or phone lines and displayed to high-resolution color monitors located in the Pathology Surgical Suites at Robert Wood Johnson University Hospitals in New Brunswick or in Hamilton. The pathologists were asked to render diagnoses using the digitized images corresponding to cases which they had interpreted up to five years earlier using traditional microscopic methods (specimens on glass slides). Clinical diagnoses rendered by pathologists using the prototypical system and software were consistent with those reported using conventional light microscopy in more than 97% of the cases studied.

In the current experiment a subset of nine digital images was selected from those which were correctly diagnosed in the first experiment. The images fell into three categories: hematopathological (images prefixed B), histopathological (images prefixed S), and stained needle biopsies extracted from liver (images prefixed C). The images were compressed with the industry standard JPEG technique [1]. Although improved performance can be achieved using modified quantization tables [11], [14], we implemented a "worst case design" approach by utilizing the standard technique.

Ten compressed versions were generated from each original (uncompressed) image. The compression was controlled by the quality factor Q of the JPEG algorithm. To establish the operational ranges of the two experiments a pilot study was run with subjects who did not participate in the core study. In the j.n.d. experiment the sequence was

$$Q = 7,\ 9,\ 12,\ 15,\ 18,\ 21,\ 25,\ 30,\ 40,\ 50$$

whereas in the diagnostic accuracy experiment

$$Q = 5,\ 6,\ 7,\ 8,\ 9,\ 10,\ 12,\ 15,\ 18,\ 20,\ 30.$$

The compression ratios were computed by dividing the total number of pixels in an image by the number of bytes in the JPEG-compressed file. The size of image header (less than a hundred bytes) can be omitted in these computations.

The achieved compression ratios varied widely. For example, Q values of 10 and 25 resulted in compression ratios in the ranges $[12.09,\ 37.63]$ and $[6.98,\ 24.29]$, respectively, across the nine different images. In Fig. 1 an example is shown for each of the three classes.

B. Procedure

Five observers participated in the experiments. Four were certified pathologists and one was a fourth-year medical student. All had normal or corrected-to-normal vision and viewed the images binocularly. Each participant was seated 45.7 cm away from the screen of a Silicon Graphics Indigo2 Extreme. The 8-bit images were shown on the 24-bit display monitor to avoid conflict in color map allocation. A 512 $\times$ 480 image covered 14.6 $\times$ 13.3 cm on the screen. None of the pathologists had any recollection of previously seeing the cases. In both experiments (j.n.d. and diagnostic accuracy) the participants were given similar instructions; the experiments differed only in the set of employed compressed images and the observer's task.

The observers were presented with two images side-by-side, the original and a compressed version, with random left/right arrangements across trials (Fig. 2). As was mentioned before, all the images employed correspond to pathological profiles consistent with routine disease states, for which diagnosis is unambiguous. The experiment was designed utilizing the "personal gold standard" motif discussed by Cosman *et al*. [3]. The personal gold standard is the most biased against image compression, and it provides the most consistent judgments. To meet the criterion of the personal gold standard the original (uncompressed) image was continuously provided.

The two-alternative forced-choice (2AFC) experimental paradigm was used. The images remained on the screen until the observer responded. The 2AFC task was to indicate whether the two images in the pair were equivalent with respect to the task: perceivable difference in the j.n.d. experiment, preserved ability to diagnose in the diagnostic accuracy experiment. The observer indicated her/his choice by clicking on one of two

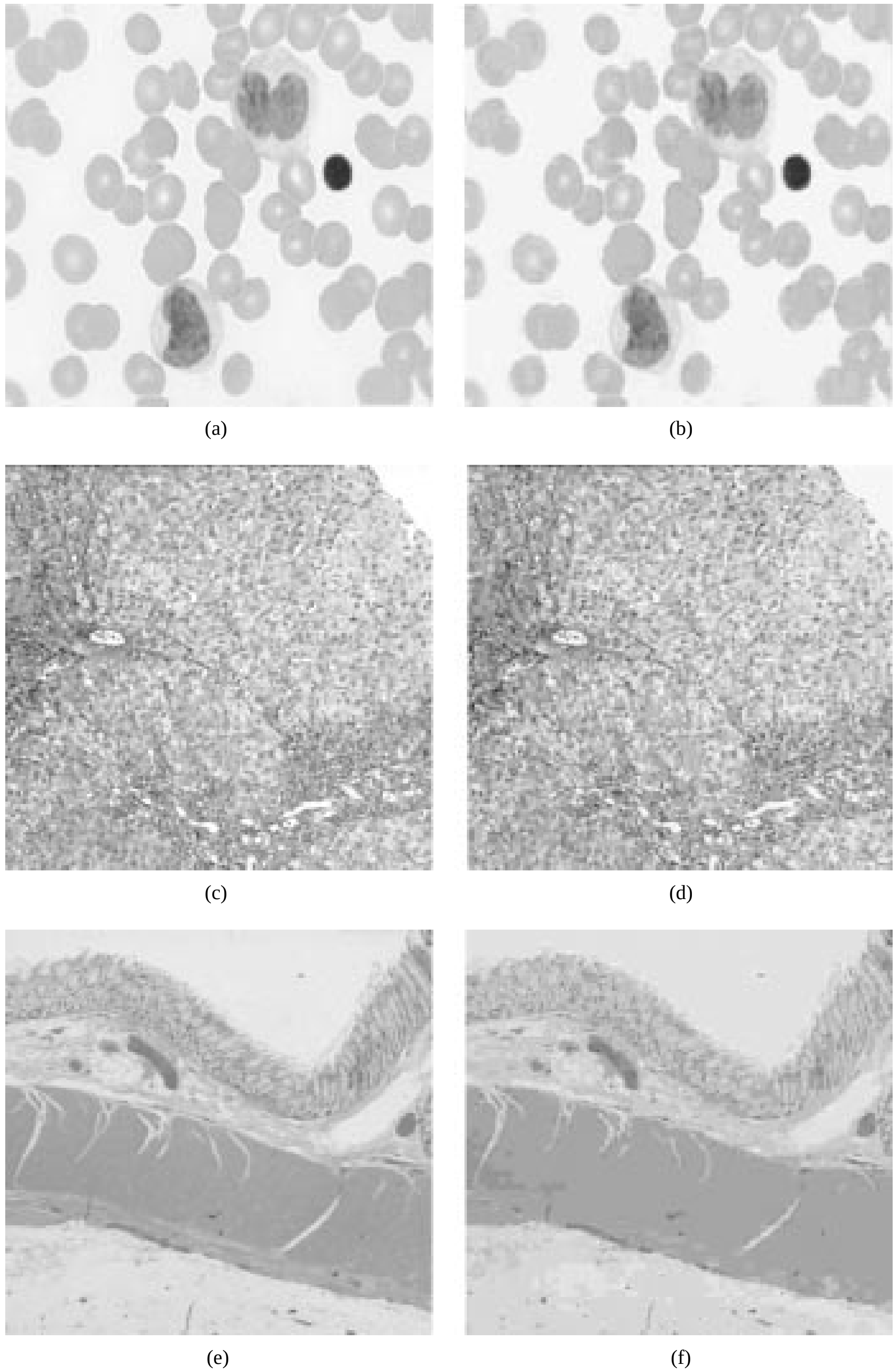


Fig. 1. Examples of images used in the tests: (a) An uncompressed hematopathological image B1. (b) Compressed at $Q = 9$, compression ratio 33.9. (c) An uncompressed histopathological image S3. (d) Compressed at $Q = 5$, compression ratio 37.4. (e) An uncompressed needle biopsy image C1. (f) Compressed at $Q = 5$, compression ratio 22.2.

bars. They could establish their own pace by delaying the response. The next trial commenced immediately after the response key was pressed. The j.n.d. and diagnostic accuracy experiments were performed in different occasions, but the data for one experiment was gathered in a single session which took about an hour.

Two staircases were employed to determine the Q value for which the observer exhibited a noticeable change in performance. One staircase proceeded from above, i.e., displaying increasingly compressed images, while the other staircase moved in the opposite direction. Once the observer indicated that there was a change detected in the image pair, a second pair was presented with the same Q number for the compressed image. If the observer again responded with a change, the staircase was terminated and the Q value was recorded.

For the descending staircase, the observer initially does not detect a difference between the original and the compressed image because the staircase starts with an almost

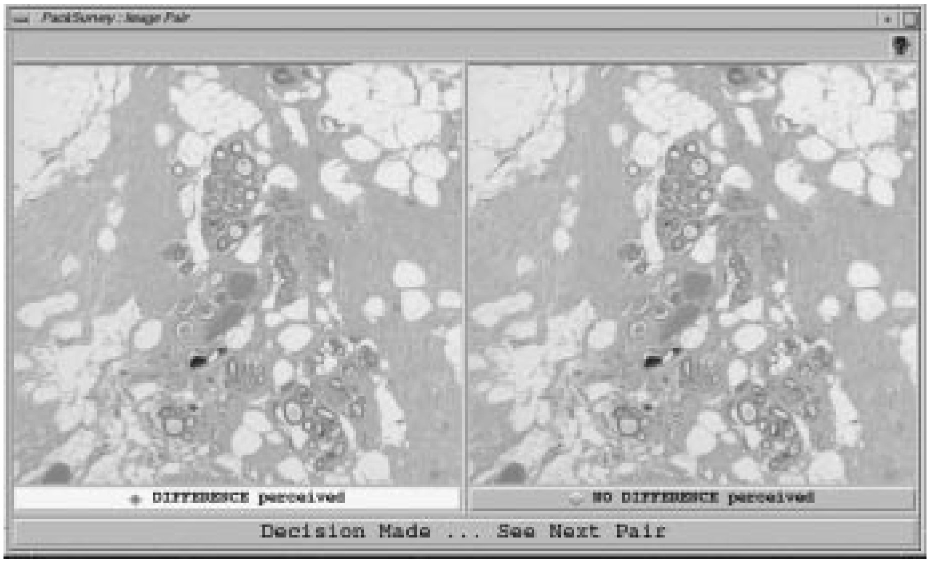


Fig. 2. Shown here is an example of stimulus (digitized breast tissue) seen in the j.n.d. experiment.

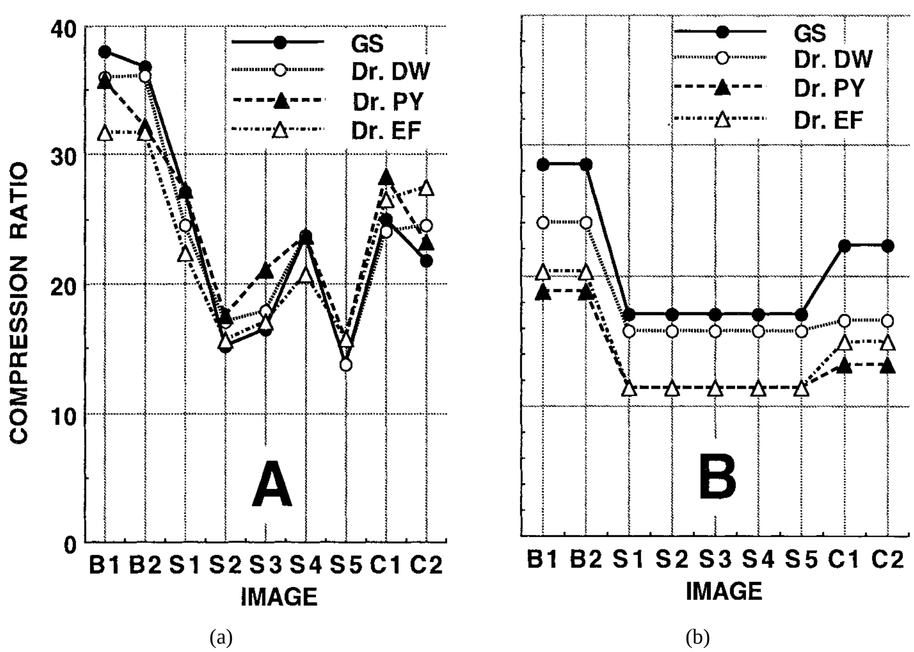


Fig. 3. Experimental results: (a) The "diagnostic accuracy" experiment. (b) The "just noticeable difference" experiment.

uncompressed image and progresses toward more compressed images as Q decreases. The goal is to obtain that value of Q, Q_{down}, which causes the observer to change his response from "No Difference Perceived" or "Diagnostic Quality" to "Difference Perceived" or "Not Diagnostic Quality." The ascending staircase starts with a very low Q, which results in a compressed image that is markedly degraded, and progresses toward less compressed images as Q increases. The goal is now to obtain that value of Q, Q_{up}, that resulted to a change in the observer's response from "Difference Perceived" or "Not Diagnostic Quality" to "No Difference Perceived" or "Diagnostic Quality." One expects the values of Q_{down} and Q_{up} to be close to each other, and this is what we obtained in practice. The average of the two, $Q^* = 0.5 * (Q_{\text{down}} + Q_{\text{up}})$ was used as the sought measure of the experiment. Values of Q below Q^* result in images that are visibly inferior to the original, given the proposed task. The Q^* values were then converted into compression ratios.

III. Results

The experimental results are shown in Fig. 3. In the "diagnostic accuracy" experiment, the same image was employed for each pair (up and down) of staircases. In the "just noticeable difference" experiment, versions having the same Q value of all the images within a category were randomly mixed across trials of a staircase. This was done to encourage the observer to look at the overall image quality, instead of concentrating on a particular feature of a single image.

High compression ratios were tolerated in the "diagnostic accuracy" experiment [Fig. 3(a)]. Note that the horizontal axis does not represent a unidimensional variable, it shows the images in a categorical order. Thus, the lines connecting the experimentally derived points do not represent a mathematical function but are included to emphasize the underlying pattern of performance as it relates to specific subspecialties and images. It is interesting to note that there is little disparity among the observers with respect to performance patterns across images. The thresholds are strongly dependent on the image category but only weakly on the specific images. The data from the "diagnostic accuracy" component experiment showed that the hematopathological images (B), histopathological images (S), and the liver biopsies (C) used in this study could be compressed by factors of about 35, 20, and 25, respectively, without compromising their clinical usefulness.

In the "just noticeable difference" experiment the observers tolerate lower compression ratios Fig. 3(b). It is important to note that for the same observer the diagnostic threshold almost always exceeded that for the j.n.d. The j.n.d component of the experiment showed that the hematopathological images (B), histopathological images (S), and the liver biopsies (C) used in this study could be compressed by factors of about 22.9, 13.9, and 16.7, respectively without decreasing subjective image quality.

IV. Conclusion

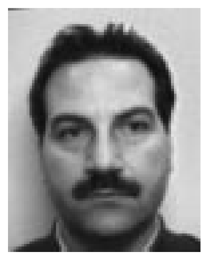
Digitized pathology presents special challenges because of the added level of complexity introduced by the chromatic content of the images. The results of this study suggest that although there is fairly large disparity among individual expert observers in their capacity to perceive distortion due to compression there is relatively good agreement among expert observers when asked to indicate how much compression is tolerable before the diagnostic essence of an image has been compromised. This study also indicates that remarkably high compression ratios are tolerable in diagnostic telepathology and that the tolerance for compression varies across subspecializations. Our results indicate that a systematic long-term study is worthwhile to investigate the minimum requirements across a much wider range of subspecializations.

Acknowledgment

The authors would like to thank K. Xu for writing the first version of the programs. The authors are also grateful to the pathologists at Robert Wood Johnson Medical School and University Hospitals who participated in the study.

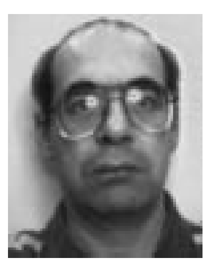
References

[1] V. Bhaskaran and K. Konstantinides, *Image and Video Compression Standards*. Boston, MA: Kluwer, 1995, ch. 5.
[2] S. Black-Schaffer and T. J. Flotte, "Current issues in telepathology," *Telemed. J.*, vol. 1, pp. 95–105, 1995.
[3] P. C. Cosman, R. M. Gray, and R. A. Olshen, "Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy," in *Proc. IEEE*, vol. 82, pp. 919–932, 1994.
[4] D. J. Foran, "A network-based prototype for interactive telemedicine and automated management of distributed clinical databases," *J. Clinical Eng.*, vol. 21, no. 5, pp. 383–392, 1996.
[5] W. E. L. Grimson, "Medical applications of image understanding," *IEEE Expert*, vol. 10, no. 5, pp. 18–28, 1995.
[6] D. J. Heeger and P. C. Teo, "A model of perceptual image fidelity," in *Int. Conf. on Image Processing*, Washington, DC, Oct. 1995, vol. II, pp. 343–345.
[7] H. K. Huang *et al.*, "Implementation of a large-scale picture archiving and communication system," *Comput. Med. Imaging Graph*, vol. 17, pp. 1–11, 1993.
[8] J. P. Kassirer, "The next transformation in the delivery of health care," *New England J. Med.*, vol. 332, pp. 52–54, 1995.
[9] E. A. Krupinski, R. S. Weinstein, K. J. Bloom, and L. S. Rozek, "Progress in telepathology: System implementation and testing," in *Advances in Pathology and Laboratory Medicine*. New York: Mosby, 1993, vol. 6, pp. 63–87, 1993.
[10] J. O. McGee, "Phone-based remote video microscopy: Telepathology's present and future," *Adv. Imag.*, vol. 10, pp. 33–35, 1995.
[11] J. Van Overloop, W. Philips, and P. De Neve, "A comparison of DCT-like transform coders for medical images," *1996 Int. Conf. on Image Processing*, Lausanne, Switzerland, Sept. 1996, vol. II, pp. 907–910.
[12] J. L. Smith, "Data compression and perceived quality," *IEEE Signal Processing Mag.*, vol. 12, no. 5, pp. 58–59, 1995.
[13] D. A. Silverstein and J. E. Farrell, "The relationship between image fidelity and image quality," *1996 Int. Conf. on Image Processing*, Lausanne, Switzerland, Sept. 1996, vol. I, pp. 881–884.
[14] B. J. Sullivan, R. Ansari, M. L. Giger, and H. MacMahon, "Effects of image preprocessing/resizing on diagnostic quality of compressed medical images," in *Int. Conf. on Image Processing*, Washington, DC, Oct. 1995, vol. II, pp. 13–16.
[15] J. M. Taylor and L. D. Picciano, "The emerging role of color imaging in telemedicine," in *Proc. IS&T's 48th Annu. Conf.*, Washington, DC, May 1995, pp. 107–110.
[16] S. Wong, L. Zaremba, D. Gooden, and H. K. Huang, "Radiologic image compression—A review" *Proc. IEEE*, vol. 83, pp. 194–219, 1995.



David J. Foran (S'89–M'91) received the B.S. degree in 1983 from Rutgers University, Piscataway, NJ, where he studied zoology and physics. He received the Ph.D. degree in biomedical engineering jointly from the University of Medicine and Dentistry of New Jersey (UMDNJ) and Rutgers University in 1992, and received postdoctoral training in computational biology at the Department of Biochemistry at UMDNJ between May 1992 and July 1993.

He has served as a physics laboratory instructor at the New Jersey Institute of Technology, Newark, and worked as a junior scientist and systems analyst at Johnson & Johnson Research, Inc., North Brunswick, NJ, from 1986 to 1988. He joined the faculty at Robert Wood Johnson Medical School in July 1993 and is currently an Assistant Professor of Pathology and Laboratory Medicine. He was recently appointed Director of the interdeparmental Center for Biomedical Imaging at Robert Wood Johnson Medical School and Robert Wood Johnson University Hospitals. He is a member of the faculty in the Program in Computational Molecular Biology and Genetics and an Assistant Professor of Research in the Department of Computer Science, both at Rutgers University. His research interests include biomedical imaging, pattern recognition, distributed systems, computer-assisted diagnosis, and medical informatics.



Peter P. Meer (S'84–M'86–SM'95) received the Dipl. Engn. degree from the Bucharest Polytechnic Institute, Bucharest, Romania, in 1971, and the D.Sc. degree from the Technion—Israel Institute of Technology, Haifa, in 1986, both in electrical engineering.

From 1971 to 1979, he was with the Computer Research Institute, Cluj, Romania, working on R&D of digital hardware. Between 1986 and 1990, he was Assistant Research Scientist at the Center for Automation Research, University of Maryland, College Park. In 1991, he joined the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ, and is currently an Associate Professor. He held visiting appointments at the Advanced Telecommunications Research (ATR) Institute, Kyoto, Japan (1992); the Korean Advanced Institute of Science and Technology, Taejon, Korea (1994); and at the Linköping University, Sweden (1996).

Dr. Meer received received an Oustanding Paper award from the journal *Pattern Recognition* in 1989, and was Cochair of the Performance vs. Methodology in Computer Vision NSF/ARPA Workshop, Seattle, 1994.

Thomas Papathomas (S'75–M'76–SM'96) was born in Kastoria, Macedonia, Greece, in 1949. He received the B.S., M.S., and Ph.D. degrees from Columbia University, New York, in 1971, 1972, and 1977, respectively, all from the Department of Electrical Engineering and Computer Science.

He was a member of the technical staff in the Electronic Power Systems Laboratory from 1977 to 1982, a supervisor of the AC Control Systems Group from 1982 to 1983, and a member of the research staff in the Visual Perception Research Department from 1983 to 1989, all at AT&T Bell Laboratories. Since 1989, he is an Associate Professor in the Department of Biomedical Engineering and the Associate Director of the Laboratory of Vision Research, Rutgers University, Piscataway, NJ. His research interests are in human and machine vision, in image processing, imaging systems, and scientific visualization techniques. He has designed several exhibits on stereopsis in science museums, and is the Editor-in-chief of *Early Vision and Beyond* (Cambridge, MA: MIT Press), a volume of interdisciplinary research in psychophysics, neurophysiology, and computational vision. Since 1993, he has served as a member of the Editorial Board of the *International Journal of Imaging Systems and Technology*.

Ivan Marsic (M'95) received the B.S. and M.S. degrees in computer engineering from University of Zagreb, Croatia, in 1982 and 1987 respectively, and the Ph.D. degree in biomedical engineering from Rutgers University, New Brunswick, NJ, in 1994.

He joined the CAIP Center, Rutgers University, in 1994. He is currently a Research Assistant Professor. His current research interests include computer networking, multimedia systems, and machine vision.