Document Systems Analysis: Testing, Testing, Testing

George Nagy
DocLab, RPI, Troy, NY, USA

Is there really a dearth of document test data, or are we merely dissatisfied with what is available? The hand digitized characters collected by W.H. Highleyman for his 1961 doctoral research may still hold the record for widest use. IBM compiled a 700,000 character database for the 200-font Optical Page Reader delivered to the Social Security Administration in 1966 (but OCR manufacturers have always considered their test data proprietary). Among the digitized data sets for sale in 1974 by the Computer Society were alphanumeric FORTRAN coding sheets, cursive script by seven writers, and 100,000 printed characters from Calspan/USPS. Beginning about 1990, large data sets were labeled at the Unipen Consortium, Concordia-in Montreal, University of Washington, University of Bern, ISRI in Las Vegas, and the National Institute of Standards and Technology, among others. ETL-8 and ETL-9 were widely used for testing Chinese and Japanese character recognition algorithms. Arabic databases were next. More specialized collections targeted signatures, tables, technical drawings, musical scores, mathematical formulas, ancient manuscripts and logos. Lacking adequate document models, OCR and even binarization error rates remain far less predictable than, for example, the course of a rocket to Mars. Therefore improvement and performance assessment of document analysis systems continues to depend on empirical experiments. What can we, and what should we, expect from the next generation of test databases?

George Nagy graduated from McGill University in Engineering Physics (fencing and chess). He earned his MS at McGill by solving Euler's Second Equation for the hysteresis motor. He was awarded the PhD at Cornell University in 1962 for helping Frank Rosenblatt build Tobermory, a sixteen-foot, four-layer neural network for speech recognition. After a short postdoc he worked on character recognition and remote sensing at IBM Yorktown (he claims credit for IBM's growth during this period). He devoted a reverse sabbatical at the Université de Montréal to recording pulse trains from cats' medial geniculate nuclei. In 1972 he joined the Department of Computer Science at the University of Nebraska where he dabbled in computational geometry, GIS and HCI. Since 1985 he has been Professor of Computer Engineering at RPI in Troy, NY. Nagy's credits in document analysis include Chinese character recognition with Dick Casey, "self-corrective" character recognition with Glen Shelton (with a reprise twenty-eight years later with Henry Baird), character recognition via cipher substitution with Casey, Sharad Seth, and Tin Ho, growing X-Y trees with Seth, table interpretation with Dave Embley, Mukkai Krishnamoorthy, Dan Lopresti and Seth, modeling random-phase noise with Prateek Sarkar and Lopresti, style-constrained classification with Sarkar, Hiromichi Fujisawa, Cheng-Lin Liu and Harsha Veeramachaneni, and recently paper-based election systems research with Lopresti and Elisa Barney Smith. In his spare time Nagy enjoys skiing, sailing, and writing prolix surveys.