2.5 Decades of Family History Technology Research at BYU

Bill Barrett
Department of Computer Science
Brigham Young University

Much of our work in Family History Technologies over the past decade has been toward the development of the *Digital Microfilm Pipeline* [1-2, Fig. 1]. This pipeline was first envisioned and architected over 10 years ago. Self-funded work on pieces of that pipeline has progressed steadily over that time, as indicated by the dates in the boxes.

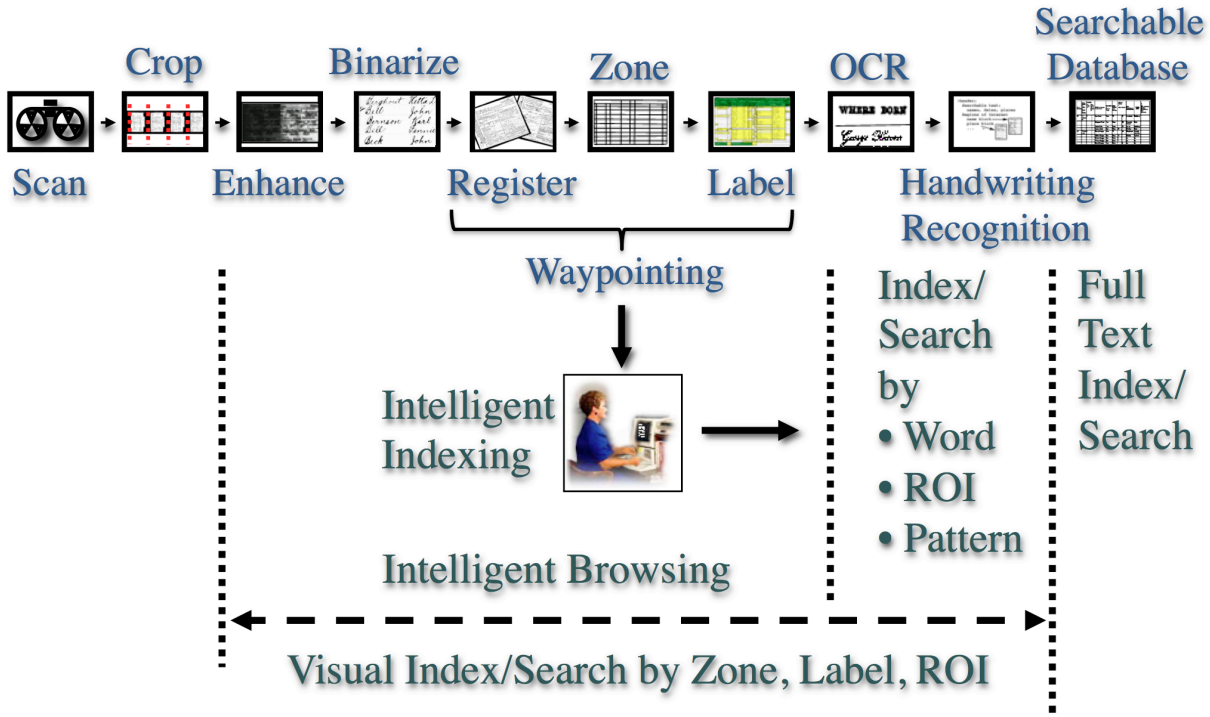

Figure 1. The Digital Microfilm Pipeline includes the tasks of scanning, cropping, deskewing, scaling, registering, image enhancement, zoning and labeling, OCRing and indexing, with companion browsing technologies for efficient, low-bandwidth internet access and information extraction.

Because we believed these technologies were important to the Church, we embarked on a path of research in Document Understanding and associated Image Analysis and Recognition methodologies, including OCR. As can easily be seen, there has been a sustained effort in the development of enabling technologies that now spans well over two decades. A short summary of theses and seminal papers in these areas follows:

1991 A One-Pass Neural Network for Probabilistic Character Recognition (Kelly Lent)
(This was the initial work on OCR using profiles)

1992 Towards Intelligent Document Understanding (Mai Zhuang)
(This was the initial work on Document Zoning – distinguishing text from graphics and images)

1993 Distributed Processing of Computationally Intensive Robot Vision Tasks (Tim Heaton)
(This was the initial work on recursive profile parsing that is now used in frame cropping)

1995 Use of Pyramidal Structures in the Generalized Hough Transform for Scale and Rotation Independent Template Matching (Yoonkoo Cho - Efficient, hierarchical search and match of arbitrary patterns)

1996 Posting Paper on the Web: Document Component Recognition and Linking (Rex Barzee)
(Paper In -> HTML Out: Automatic Recognition of Document Components (text, captions, graphics), highly accurate OCR engine, automatic link generation)

1998 Breakpoint Skeletal Representation and Compression of Document Images (Kai Wing Tam)
(Medial Axis compression and representation of scanned characters)

1999 Automatic Cost-Ordered Edge Linking (Glen Sawyer)
(Robust linking of lines in noisy images, such as handwriting)

2001 Houghing the Hough: Peak Collection for Detection of Corners, Junctions, and Line Intersections
(Kevin Petersen – Robust (straight) line detection, such as in document tables)

2003 Consensus-Based Table Form Recogntion (Heath Nielson)
(Automatic document zoning, zone labeling, and creation of document template)

2003 Just-In-Time Browsing for Digital Images (Doug Kennard)
(Completely scalable browsing technology – browse documents over low-bandwidth modem)

2003 Fast Registration of Tabular Document Images Using the Fourier-Mellin Transfrom (Luke Hutchison)
(Automatic detection and reversal of document transformations, including scale, skew and rotation)

2004 Decision Tree Discrimination of Text Types (Sam Pinson)
(Automatic discrimination of different types of text, especially machine print vs. handwriting)

2004 Automatic Indexing of Ribbon Microfilm (Mark Pinson)
(Automatic indexing of the content of a scanned ribbon of microfilm – from frame cropping (macro level) to zone indexing (micro level))

2011 Recursive Otsu Thresholding Method for Scanned Document Binarization. Treats handwriting as multi-layered, consisting of dark strokes, light strokes and in between. Recursive thresholding peals away layer of strokes and composites them together for the complete binarized result.

2011 Word Warping for Offline Handwriting Recognition. This is a break-through technology that recognizes handwriting based on morphing one piece of handwriting to another and then computing the distance, or difference, between the two glyphs. The less the stretch/distance, the closer the match. In-vocabulary recognition rates are approaching 90% with 95% in the top three.

2011 Connected Component Level Discrimination of Handwritten and Machine-Printed Text Using Eigenfaces. A Principle Component Analysis technique for discriminating between machine print and handwriting.

2011 Linking the Past: Discovering Historical Social Networks from Documents and Linking to a Genealogical Database. Using the FamilySearch Database and Historical indexes to discover historical social networks (HSNs). Which people knew each other? And wrote about each other? What was their HSN? HSNs can be used to infer histories about people who never kept them based on what others have written.

2012 Further developments of word warping for handwriting recognition, including applications to other languages (French, Dutch, Chinese) and application to forgery detection.

2012 Text extraction and noise removal from Cemetery Headstones.

2015 Intelligent Indexing: A Semi-Automated, Trainable System for Field Labeling. Leverages word-warping handwriting recognition for semi-automated labeling and simultaneous transcribing of multiple fields in tabular documents.

2015 Segmentation of Cursive Handwriting in Tabular Documents for segmenting and disentangling overlapping handwritten strokes.

2016 Intelligent Pen: A Least Cost Search Approach to Stroke Extraction in Historical Documents. Use of Intelligent Scissors technology to trace strokes in cursive handwriting. A Preprocessing step for handwriting analysis and recognition.

2016 Major breakthrough in handwriting recognition using a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM) network (READNET1.BYU) for recognition of medieval (1400's – 1700's) German Handwriting. 5% Character Error Rate (CER), ~19.7% Word Error Rate (WER).

2016 Development of Subword Spotting technology for semi-automated transcription and performing subword recognition for document searching and simi-automated indexing.

2016 Creation of a system for extraction of genealogical information from Jia Pu (Chinese Genealogical Records) and for creating training sets for recognition of Chinese characters and words.

2017 Symbol Stitching for Handwriting Recognition – recombine handwritten characters to create novel words and increase training data for handwriting recognition.

2017 Use of Data Augmentation to produce state-of-the-art results (< 2.82% CER, 4.97% WER on the IAM dataset of 600 different authors) and (< 1.36% CER, 2.85% WER on the RIMES dataset), eclipsing previous best results by other researchers.

2017 First place in the ICDAR 2017 competition for reading handwriting in old (19^{th} century) letters in German, French and English. 10,000 different letters, different authors, languages. ~7% CER, 17% WER. Next best was 25% CER, 42% WER.

2017 Creation of a Neural Network for Semantic Labeling of Document Components (Machine Print, Handwriting, Lines, Stamps)

2017 First place in the pixel labeling competition for ICDAR/HIP.

2017 PageNet – a Neural Network for segmenting the top page of an opened book.

Papers and Competitions

William Barrett and Rex Barzee: "Posting Paper on the Web," Proceedings Vision Interface '98, pp. 381-388**.**

David Tam, William Barrett, Bryan Morse and Eric Mortensen: "Breakpoint Skeletal Representation and Compression of Document Images," IEEE Data Compression Conference (DCC '98), pp. 75, Snowbird, Utah, March 1998.

D. J. Kennard and W.A. Barrett: "Just-In-Time Browsing for Digital Images," Accepted for publication in IEEE Data Compression Conference (DCC 2001), Snowbird, Utah, March 2001.

D. J. Kennard and W. A. Barrett, "Just-In-Time Browsing for Digitized Microfilm and Other Similar Image Collections, *7th International Conference on Document Analysis and Recognition* (*ICDAR*), pp. 233-237, Edinburgh, Scotland, August, 2003.

H. Nielson and W. A. Barrett, "Consensus-Based Table Form Recognition, *7th International Conference on Document Analysis and Recognition* (*ICDAR*), pp. 906-910, Edinburgh, Scotland, August, 2003.

W. A. Barrett, L. Hutchison, D. Quass, H. Nielson, and D. Kennard, "Digital Mountain: From Granite Archive to Global Access," *IEEE Proceedings, International Workshop on Document Image Analysis for Libraries* (*DIAL 2004*), pp. 104-121, Palo Alto, CA, January, 2004.

L. Hutchison and W. A. Barrett, "Fast Registration of Tabular Document Images Using Fourier-Mellin Transform," *IEEE Proceedings, International Workshop on Document Image Analysis for Libraries* (*DIAL 2004*), pp. 253-269, Palo Alto, CA, January, 2004.

Douglas J. Kennard and William A. Barrett, "Separating Lines of Text in Free-Form Hand- written Historical Documents," IEEE Proceedings, 2nd International Conference on Document Image Analysis for Libraries (DIAL 2006), pp. 12-23, Lyon, France, April, 2006.

Elisa H. Barney Smith, Henry Baird, William Barrett, Frank Le Bourgeois, Xiaofan Lin, George Nagy, and Steve Simske, "DIAL 2004 Working Group Report on Acquisition Quality Control," IEEE Proceedings, 2nd International Conference on Document Image Analysis for Libraries (DIAL 2006), pp. 373-376, Lyon, France, April, 2006.

Luke A. D. Hutchison and W. Barrett: "Fourier–Mellin registration of line-delineated tabular document images," Volume 8, Numbers 2-3, *International Journal on Document Analysis and Recognition*, June 2006, pp. 87 - 110.

H. Nielson and W. Barrett: "Consensus-based table form recognition of low-quality historical documents," Volume 8, Numbers 2-3, *International Journal on Document Analysis and Recognition*, June 2006, pp. 183 – 200.

Douglas J. Kennard and W. Barrett: "Interactive Training for Handwriting Recognition in Historical Document Collections,", *Document Recogntion and Retrieval XIV*, January 2007, Palo Alto, CA.

O. Nina, B. Morse, and W. Barrett. A Recursive Otsu Thresholding Method for Scanned Document Binarization. In IEEE Workshop on Applications of Computer Vision (WACV) 2011, Kona, Hawaii, January 5-6, 2011.

Douglas J. Kennard, William A. Barrett, and Thomas W. Sederberg. Word Warping for Offline Handwriting Recognition. In 11th International Conference on Document Analysis and Recognition (ICDAR2011), volume 1, pages 1349-1353, Beijing, China, September, 2011.

Samuel J. Pinson and William A. Barrett. Connected Component Level Discrimination of Handwritten and Machine-Printed Text Using Eigenfaces. In 11th International Conference on Document Analysis and Recognition (ICDAR2011), volume 1, pages 1394-1398, Beijing, China, September, 2011.

Douglas J. Kennard, Andrew M. Kent and William A. Barrett. Linking the Past: Discovering Historical Social

Networks from Documents and Linking to a Genealogical Database. First International Workshop on Historical Document Imaging and Processing (HIP'11), volume 1, pages 43-50, Beijing, China, September, 2011.

Douglas J. Kennard, William A. Barrett and Thomas Sederberg. Handwriting Recognition (HR) of Family History Documents using a 2-D Warping-based Word-level HR Approach. Family History Technology Workshop 2012 (FHTW2012@Rootstech) (http://fht.byu.edu/prev_workshops/workshop12/), pages 50-53, Salt Lake City, February 3, 2012.

Cameron Christiansen and William A. Barrett. Removing the Noise from Cemetery Headstones. Family History Technology Workshop 2012 (FHTW2012@Rootstech) (http://fht.byu.edu/prev_workshops/workshop12/), pages 66-71, Salt Lake City, February 3, 2012.

Robert Clawson and William A. Barrett. Extraction of Handwriting in Tabular Document Images. Family History Technology Workshop 2012 (FHTW2012@Rootstech) (http://fht.byu.edu/prev_workshops/workshop12/), pages 76-79, Salt Lake City, February 3, 2012.

Douglas J. Kennard and William A. Barrett. Automatic "Life Sketch" Videos. Family History Technology Workshop 2012 (FHTW2012@Rootstech) (http://fht.byu.edu/prev_workshops/workshop12/), pages 117-119, Salt Lake City, February 3, 2012.

Douglas J. Kennard, William A. Barrett, and Thomas W. Sederberg, "Offline Signature Verification and Forgery Detection Using a 2-D Geometric Warping Approach," in 2012 International Conference on Pattern Recognition (ICPR 2012), pp. 3733-3736, Tsukuba, Japan, Nov 2012.

Cameron Christiansen and William A. Barrett. Data Acquisition from Cemetery Headstones. Document Recognition and Retrieval XX (DRR 2013) [8658-16], San Francisco, February 5-7, 2013.

Robert Clawson, Kevin Bauer, Glen Chidester, Milan Pohontsch, Douglas Kennard and William Barrett. Automated recognition and extraction of tabular fields for the indexing of census records. Document Recognition and Retrieval XX (DRR 2013) [8658-17], San Francisco, February 5-7, 2013.

Glen Chidester and William A. Barrett. Glyph Matching for Text Extraction on Headstones. Family History Technology Workshop 2013 (FHTW2013@Rootstech) (http://fht.byu.edu/), Salt Lake City, March 22, 2013.

Douglas J. Kennard, William A. Barrett, and Thomas W. Sederberg. Many-Author Offline Handwriting Recognition Using a Warping-Based Approach. Family History Technology Workshop 2013 (FHTW2013@Rootstech) (http://fht.byu.edu/), Salt Lake City, March 22, 2013.

Robert Clawson and William Barrett. Green ICR: Semi-Automated Census Record Indexing with Emphasis on Human Computer Interaction. Family History Technology Workshop 2013 (FHTW2013@Rootstech) (http://fht.byu.edu/), Salt Lake City, March 22, 2013.

Kevin Bauer and William Barrett. Better Historical Document Indexing Using Waypointing and ROI Data. Family History Technology Workshop 2013 (FHTW2013@Rootstech) (http://fht.byu.edu/), Salt Lake City, March 22, 2013.

Robert Clawson and William Barrett. Intelligent Indexing: A Semi-Automated, Trainable System for Field Labeling. Family History Technology Workshop 2014 (FHTW2014) (http://fht.byu.edu/), March 20, 2014.

Kevin Bauer and William Barrett. Intelligent Pen: A Least-Cost Search for Tracing of Handwriting. Family History Technology Workshop 2014 (FHTW2014) (http://fht.byu.edu/), March 20, 2014.

Curtis Wigington, Ryan Cheatham and William Barrett. Virtual Pedigree - Genealogy Without Borders. Family History Technology Workshop 2014 (FHTW2014) (http://fht.byu.edu/), March 20, 2014.

Robert Clawson and William Barrett. Intelligent Indexing: A Semi-Automated, Trainable System for Field Labeling. Accepted for publication/presentation at Document Recognition and Retrieval XXII (DRR 2015), Feb. 11, 2015. *Best Student Paper*

Brian Davis, William Barrett and Scott Swingle. Segmentation of Cursive Handwriting in Tabular Documents. Accepted for publication/presentation at Document Recognition and Retrieval XXII (DRR 2015), Feb. 11, 2015.

Kevin Bauer and William Barrett. Intelligent Pen: A Least-Cost Search for Tracing of Handwriting. Family History Technology Workshop 2015 (FHTW2015) (http://fht.byu.edu/), February 10, 2015.

Brian Davis, Robert Clawson and William Barrett. Subword Spotting for Use in a Computer Assisted Transcription System. Document Analysis Systems 2016 (DAS2016), Santorini, Greece, April 11-14, 2016.

Kevin L Bauer and William Barrett. Intelligent Pen: A Least Cost Search Approach to Stroke Extraction in Historical Documents. Document Recognition and Retrieval XXIII (DRR-057.1-057.10, 2016), Feb. 17, 2016. *Best Student Paper*

Brian Davis, Robert Clawson and William Barrett. Flexible Computer Assisted Transcription of Historical Documents Through Subword Spotting. Family History Technology Workshop 2016 (FHTW2016) (http://fht.byu.edu/), February 2, 2016.

Curtis Wigington and William Barrett. Improved Consensus Path in Intelligent Pen. Family History Technology Workshop 2016 (FHTW2016) (http://fht.byu.edu/), February 2, 2016.

Curtis Wigington and William Barrett. Geometric Distortion for Handwriting Recognition. Family History Technology Workshop 2017 (FHTW2017) (https://fhtw.byu.edu/program), February 7, 2017.

Seth Stewart and William Barrett. Symbol Stitching for Handwriting Recognition. Family History Technology Workshop 2017 (FHTW2017) (https://fhtw.byu.edu/program), February 7, 2017.

Maroua Mehri, Pierre Heroux, Remy Mullot, Jean-Philippe Moreux, Bertrand Couasnon and Bill Barrett. HBA 1.0: A Pixel-based Annotated Dataset for Historical Book Analysis. Fourth International Workshop on Historical Document Imaging and Processing (HIP'17), volume 1, pages 107-112, Kyoto, Japan, November 10, 2017.

Seth Stewart and Bill Barrett. Document Image Page Segmentation and Character Recognition as Semantic Segmentation. Fourth International Workshop on Historical Document Imaging and Processing (HIP'17), volume 1, pages 101-106, Kyoto, Japan, November 10, 2017.

Chris Tensmeyer, Brian Davis, Curtis Wigington and William Barrett. PageNet: Page Boundary Extraction in Historical Handwritten Documents. Fourth International Workshop on Historical Document Imaging and Processing (HIP'17), volume 1, pages 59-64, Kyoto, Japan, November 10, 2017.

Curtis Wigington, Seth Stewart, Brian Davis, Bill Barrett, Brian Price, and Scott Cohen. Data Augmentation for Recognition of Handwritten Words and Lines Using a CNN-LSTM Network. In 14th International Conference on Document Analysis and Recognition (ICDAR 2017), volume 1, pages 639-645, Kyoto, Japan, November, 2017.