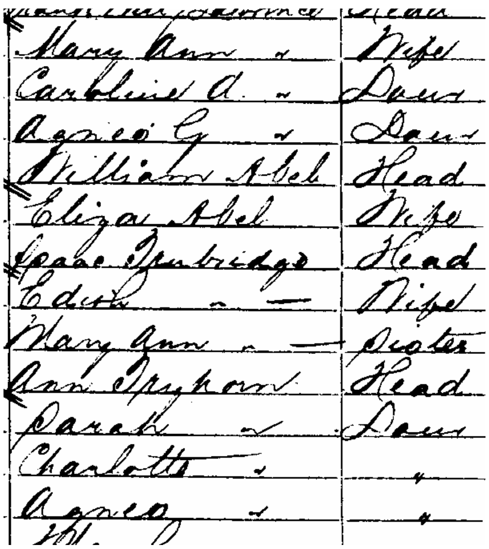
EXTENDED ABSTRACT

Handwriting Recognition for Genealogical Records

Luke Hutchison (lukeh@email.byu.edu); Thesis advisor: Dr. Tom Sederberg, C.S.Dept.

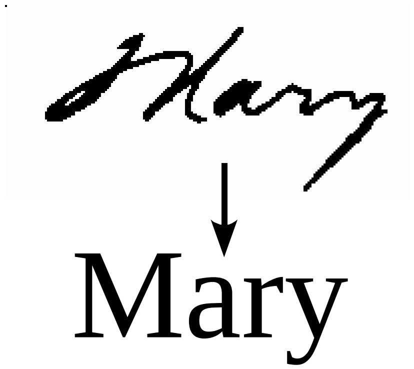
The Church of Jesus Christ of Latter-day Saints currently has one and a half billion pages of handwritten genealogical records on microfilm, with an additional fifty thousand pages being added each month. Locating a specific name amongst these records usually takes a great deal of time and patience. Much of the handwriting is archaic and difficult to read.


Mary Ann
Wife
Caroline A.
Daur
Agnes G
Daur
William Abel
Head
Eliza Abel
Wife
Isaac Trowbridge
Head
Edith
Wife
Mary Ann
Sister
Ann Seyhorn
Head
Sarah
Daur
Charlotte
Agnes

An automatic handwriting recognition system is being developed at BYU which will allow names to be read and extracted from microfilms automatically by a computer, and entered into a searchable database. This will allow a genealogical researcher to enter a name they are looking for into a search engine, and have a list of possible matches be presented to them for examination. The user will be able to select any of the promising results, and be taken to an image of the full microfilm page corresponding to that result.

Reliable automatic handwriting recognition has always been a difficult task, even for well-formed words written with modern writing instruments. Older genealogical records are of course all handwritten, usually with some sort of quill pen, and many records have faded with time or become damaged in other ways. Problems often faced in trying to recognize handwriting from old records then involve ignoring some superfluous information that is present (e.g. ink blobs), and inferring information that is not present (e.g. when a stroke is broken due to fading). One advantage to older handwriting is that it is reasonably consistent from line to line for a single author, as scribes tended to be meticulous in their penmanship.

To improve the accuracy of such a system, domain-specific knowledge must be taken into account. For example, if a particular character is ambiguous, then recognition accuracy of it can be increased by (a) looking at the possible readings of the letters before and after, and consulting a table of letter-pair frequencies in names; and (b) Referring to an extensive name-list, each of which is annotated with a frequency of occurrence for the relevant time period. Both of these reference databases will need to be different from country to country, and from language to language. Using these references, the handwriting recognizer will be able to adjust its interpretation of particular shapes, and make informed judgment-calls in much the same way that a human might do. Ultimately multiple handwriting styles will need to be handled too.



This handwriting recognition project is just part of the overall system. Genealogical tables, for example pages of census reports, contain a lot of additional information, such as birth date, birth place and profession. Other BYU students have developed or are developing software to enhance the images; divide each page up into table cells by looking for straight lines; read the typed column headings using “OCR” (Optical Character Recognition, different from handwriting recognition); create an ontology (or conceptual map) of how each column or cell relates to others in meaning; pass the ontology and the graphical contents of each cell to the handwriting-recognizer; and then facilitate browsing the results efficiently online over a low-speed Internet connection.

Some of the difficulties inherent in building a reliable handwriting recognition system are discussed in this presentation, along with current research directions being taken at BYU to try to solve some of these problems and to produce a useful working system as a tool for genealogical use.