Corpus and Evaluation of Handwriting Recognition of Historical Genealogical Records

Patrick Schone[§], Heath Nielson[§], Mark Ward[‡]
[§]FamilySearch and [‡]LDS Publishing Services Department
50 E North Temple, Salt Lake City, Utah 84150
E-mail: {patrickjohn.schone, nielsonhe, wardrm}@ldschurch.org

Abstract

Over the last few decades, significant strides have been made in handwriting recognition (HR), which is the automatic transcription of handwritten documents. HR often focuses on modern handwritten material, but in the electronic age, the volume of handwritten material is rapidly declining. However, we believe HR is on the verge of having major application to historical record collections. In recent years, archives and genealogical organizations have conducted huge campaigns to transcribe valuable historical record content with such transcription being largely done through human-intensive labor. HR has the potential of revolutionizing these transcription endeavors. To test the hypothesis that this technology is close to applicability, and to provide a testbed for reducing any accuracy gaps, we have developed an evaluation paradigm for historical record handwriting recognition. We created a huge test corpus consisting of four historical data collections of four differing genres and three languages. In this paper, we provide the details of these extensive resources which we intend to release to the research community for further study. Since several research organizations have already participated in this evaluation, we also show initial results and comparisons to human levels of performance.

Keywords: Handwriting recognition, historical image processing, multilinguality

1 Overview

Censuses, birth records, and other types of historical record collections have significant value for genealogical and related kinds of research. These records are crucial for identifying people and key life events, so archival and genealogical organizations have expended significant efforts to obtain and transcribe them. Most of this transcription has been done through massive, crowd-sourced, human efforts. Human transcription processes are beneficial in that they can often provide high levels of accuracy. Yet such methods are expensive, if the workforce is paid; or it is limited by the availability, skills, frustration tolerance, and retention of volunteer annotators. Automatic transcription of images to either supplement or replace human labor would therefore have significant value.

Over especially the last decade, there have been major advances in automatic transcription of offline (i.e., previously-written) handwritten documents. This automatic transcription is often referred to as *handwriting recognition* (HR). The DARPA (2008) MADCAT program and the NIST OpenHaRT (NIST, 2010) evaluation, for examples, have allowed the research community to focus HR efforts on modern non-English documents and systems have now reached respectable levels of accuracy for some document types. Could such technologies be re-applied to genealogically-relevant documents? Moreover, if a handwriting corpus of genealogical information were released to the research community, would it foster further improvements?

One organization with particular interest in determining answers to these questions is FamilySearch. According to Wikipedia, FamilySearch "is a genealogy organization operated by The Church of Jesus Christ of Latter-day Saints. It is the largest genealogy organization in the world." (Wikipedia:FamilySearch, 2014) FamilySearch has access to *billions* of historical images which are very beneficial for genealogical research. FamilySearch uses volunteers to transcribe its image collections and is able to create hundreds of millions of searchable records each year. Even so, the influx of new records often outpaces the rate of transcription. Moreover, the availability of volunteers in non-English languages is limited. Thus, the need for automation becomes paramount.

We have created an evaluation called *IRIS*, using FamilySearch documents, which consists of (a) an evaluation framework; (b) evaluation tools and methods; and most importantly, (c) a huge collection containing almost 50000 transcribed historical images spanning three different languages and four collection types. In addition, IRIS has evaluated human performance for the major English collection with the purpose of making human-to-automation comparisons. The IRIS evaluation tools allow systems to be scored with a weighted word error rate (WWER), which favors information of higher genealogical value (such as personal name components).

A number of major research organizations with HR systems have already participated in IRIS evaluations. The best results for their systems can be taken as baselines for future researchers who use the IRIS collection, and will therefore be included in this paper. (However, it should be mentioned that for the original IRIS participants, the evaluation set was completely blind, and they were only given two weeks to apply their algorithms to the held-out test set).

Given results from OpenHaRT and experience with human error rates coupled with some of the particular

issues with the IRIS collection data, we initially expected weighted word error rates (WWERs) to exceed 40% in English and worse in other languages. Much to our surprise, systems were able to achieve WWERs as low as 19.6%! Moreover, this evaluation was able to note that HR systems can get extremely high performance levels for small-vocabulary tasks. We believe that outcomes thus far suggest that HR systems are ripe for providing significant benefit to genealogical and historical record transcription. We therefore wish to make IRIS available to the community to see if researchers can span the gap of the few percentage points of error between humans and automation and thus make HR highly applicable to these kinds of documents.

In this paper, we describe this extensive evaluation more fully. We provide documentation about the collections and metrics, and we show the best results on each collection as contrasted with some human performance.

2 Handwriting Recognition Background

Handwriting recognition (HR), at least in some form, has been studied for almost a century (Goldburg, 1914; Hansel, 1939). Yet large-scale common collections of offline HR have really emerged within the last twenty years; and major evaluations of offline handwriting have been of still more recent advent. "Offline" recognition, in these cases, means that a system is asked to transcribe the handwriting in some document after the scribing has already occurred (as opposed to *online* which transcribes real-time as a person is writing).

A seminal evaluation paradigm of offline HR was created for Arabic documents in a partnership between the Linguistic Data Consortium (LDC) and the U.S. National Institute of Standards and Technology (NIST), and in association with the DARPA MADCAT program. This evaluation paradigm consisted of a corpus-creation phase followed by an international evaluation. For the corpus-creation, since Arabic handwritten documents were scarce, and given that the LDC already had access to many parallel text corpora, Strassel (2009) gathered native Arabic participants to write down by hand what was already written in texts. Scribes were asked to use various writing implements, and the results of their scribing were scanned at 600 DPI to create a handwriting recognition corpus. The output of this work then served as a wide-scale training and evaluation corpora for MADCAT and for NIST's OpenHaRT evaluation, both of which were mentioned earlier. At OpenHaRT 2010, several systems participated in the HR portion of the evaluation, and achieved word error rates (WER) as low as 37.7% on word-segmented handwritten documents.

A handwritten corpus of Chinese documents has since been created by the LDC in a similar fashion (Song, et al., 2012). There has also been another OpenHaRT (NIST, 2013) evaluation (which postdates our corpus creation and initial evaluations) with systems yielding much better results on even line-segmented documents.

Many other offline handwriting databases are available. These have typically not been used for major evaluations, but some are used for cross-comparisons. A commonly-used collection is the George Washington papers collection (Lengel, 2008). Other HR databases non-exhaustively include collections in English (Marti and Bunke, 2002); Spanish (España, et al., 2004); French (Viard-Gaudin, et al., 1999); and other languages like Chinese (Su, et al.,2007, Liu., et al., 2011).

3 IRIS Evaluation Collections

We sought to determine handwriting recognition's applicability to the wide range of *genealogically*-valuable collections that have been generated over the centuries. For such a test, language variability would be valuable, as would be the ability for a system to apply to form-based tables, fill-in-the-blank documents, and free-form writings. IRIS consists of four different data sets which were selected to study system performance along these multiple dimensions. The collections will be described in section 3.2, but their names and the number of images per collection are provided in Table 1. It should be mentioned that whereas other HR collections typically have from one to hundreds of scribes, these collections are based on the writings of thousands of different census-takers, ecclesiastical leaders, and court officials.

Corpus	Training Size	Evaluation Size
1930 US Census	15,061	1,673
1930 Mexico Census	8,652	961
Arkansas Marriages	7,502	834
French Parish Records	10,529	1,170

Table 1: Numbers of Images Per Collection of IRIS

3.1 Description of "Gold Standard"

The IRIS training transcripts were created by hosts of volunteer annotators. FamilySearch has a crowd-sourcing infrastructure called FamilySearch Indexing (FamilySearch, 2014) whereby a willing participant can provide transcriptions for any of a number of different historical collections. These volunteers possess a wide-range of skills, so FamilySearch sends each image that needs transcription to two independent annotators. Any errors are then adjudicated by an arbitrator.

Though these transcripts are doubly transcribed and reviewed, they still contain errors. Moreover, the transcriptions were prepared for purposes independent of IRIS, so there are conventions that were followed which resulted in non-verbatim transcriptions. For example, a person in a census who was born in Pennsylvania may have a birth place listed as Pennsylvania, PA, Penn, etc., or by some ditto information (DO or ‘’) but likely are transcribed as "Pennsylvania." In addition to these inexactness issues, certain non-genealogically-relevant phrases and fields were not transcribed. These phenomena require special attention during system building but they have also been taken into consideration in the scoring tools. This will be described later.

In the case of the 1930 U.S. Census, the *evaluation* data was re-transcribed by a commercial entity with a target level of performance of at least 99.5% accuracy. One-tenth of the results that were provided through this effort were subsequently reviewed by two separate, highly skilled volunteers who carefully studied the transcriptions and raw data and identified any anomalies. This vetting process estimated accuracy at 99.7%. This high accuracy was desirable to ensure that any observed errors for at least one of the IRIS collections would be almost exclusively due to HR systems.

The last confounding issue for system-builders for the data sets is that IRIS provides no bounding boxes around the key genealogical facts on each page. So systems need to attempt to automatically detect the layout of each page, and identify the columns, rows, and boxes. Note that this is a significant departure from what is provided in evaluations like OpenHaRT where participants are given word or line segmentation boundaries.

3.2 Image Collection Descriptions

3.2.1 United States 1930 Census

Figure 1: United States 1930 Census

The 1930 US census ("United States Census, 1930", 2014) records, as depicted in Figure 1, are tabular forms. Most forms contain 50 rows, but not all rows need to be completed. Each row represents statistical information about a particular individual, including his or her name, gender, age, race, origin, parental origin, and association with head of household. Some of these entries, such as race and gender, will have information that is drawn from small vocabularies. Other fields, like places of origin and personal names, involve extremely large vocabularies since they touch upon every US person and locations that existed in the 1930s.

These US census records were selected for IRIS because they are in English, because they are tabular, and are considered to be some of the most genealogically-beneficial document collections. Moreover, since volunteer patrons enjoy transcribing censuses because handwriting is fairly clear, this suggests that automation may also be easiest for these collections. Lastly, data from the 1920 and earlier US censuses exists, so system-builders could conceivably leverage previous collections for language modeling. At the time of the IRIS data preparation, the 1930 Census was the most-recently available census and was only partially completed. For IRIS, information from 35 US states are represented in the collection.

For this particular task, as with other IRIS census-transcribing tasks, systems are required to identify the cells of information within their appropriate rows and columns, and then provide automatic transcription for headings and particular columns of interests. No bounding boxes are provided to system builders.

3.2.2 Mexico 1930 Census

Figure 2: 1930 Mexico Census

For HR to have maximal value, it must be reconfigurable to new languages with limited effort. To test language portability, the Mexico 1930 Census ("México, Censo Nacional, 1930", 2014) seen in Figure 2, which is completely in Spanish, was also chosen for use in IRIS. Like the US Census, the Mexico census consists of 50 rows each focused on a particular individual; and it likewise has the benefit of being an extremely rich collection of genealogical facts.

However, this collection also has a feature that was not observed in the US Census. That is, this collection has the interesting property that marriage information is conveyed using columnar checkboxes. The census-taker was instructed to mark an "X" in the *Soltero* column if the individual was single, an "X" in the *Casado Por Lo Civil* column for civilly married, etc. Therefore HR systems need to identify the correct data column in order to properly transcribe the information.

Another difference for this collection is that, to the best of our knowledge, no other Mexico censuses have been transcribed nor released. Thus, this census serves as a collection for researchers to study HR performance in the absence of existing external language modeling material.

3.2.3 Arkansas Marriage Collection

Many genealogical collections are fill-in-the-blank templates. In such cases, there may be many words on the

page which are part of the pre-printed form, and the record-keeper's job was to fill specific empty fields in the page with vital information. The Arkansas Marriage Collection ("Arkansas Marriages, 1837-1944",2014), seen in Figure 3, is an example of such a collection. Whereas censuses are typically only taken once or twice a decade, these vital records are generated daily and thus represent key genealogical repositories.

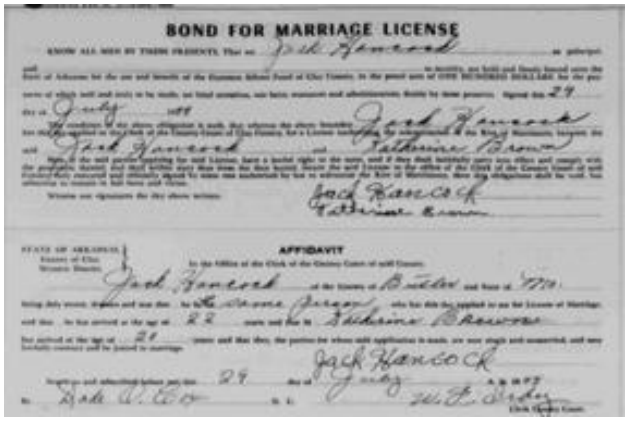

Figure 3: Arkansas Marriages

In terms of research benefits, a collection such as this helps to determine how well systems can (a) identify the locations of the slot-filling areas are on a page, and (b) mine the names, dates, and places which are spread throughout that page. The fact that vital information is not co-located in the page makes transcription somewhat more difficult than censuses for humans, but it can have added benefits for automation. As a specific example, in this collection, personal names may appear in multiple places in each image. In Figure 3, the groom's name, "Jack Hancock" appears four times throughout the image. System-builders can use such repeats to increase the accuracy of their system hypotheses. On the other hand, ascenders and descenders from the scribing (such as the lower loop in the letter "J" of "Jack") can intersect and get confused with the pre-printed form.

3.2.4 French Parish Records

Before there were censuses or pre-printed forms, one of the main methods of preservation of vital records concerning individuals came from the logbooks of ecclesiastical officials. For example, if people were born, buried, or married in their parish, dutiful priests would record such information. Church log books have existed for centuries – back at least to the 1500s. However, records of this type are extreme challeges for HR. French parish records ("France, Diocèse de Coutances et d'Avranche, registres paroissiaux, 1533-1906", 2014) are no exception. Figure 4 depicts a christening record from one on the French parishes.

If every word of the document were transcribed, this collection would be similar in nature to other collections (such as OpenHaRT). However, in IRIS, system developers are only provided with the transcription of vital facts from the page. This task therefore becomes one of trying to determine what vocabulary information repeats from document to document, and what are the novel vital facts which need to be extracted from the page.

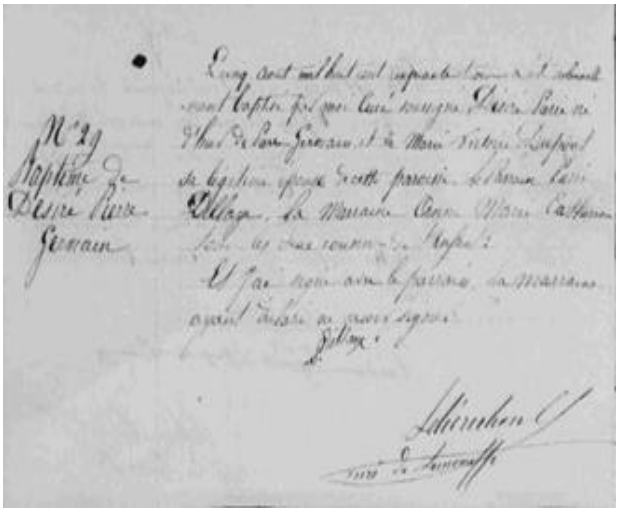

Figure 4: French Parish Records

4 Preparing for Evaluation Scoring

After having described the four multilingual collections that are involved in IRIS, we next describe methodologies for evaluating system performance.

4.1 IRIS Metrics

Word error rate (WER) is a common metric used for the automatic transcription of media. WER computes the ratio given by the sum of insertion, substitution, and insertion errors divided by the number of words in a correct transcript. With WER, every word is equally valued, so the insertion of a word like "the" counts the same as the omission of a word like "Richardson." (100%-WER), or *accuracy*, was used in OpenHaRT.

Genealogical words are not equally valuable. Mistranscription of gender information, for example, is genealogically less critical than making errors on personal names (like "Samuel") or locations (such as "Boston"). Consequently, IRIS uses *Weighted Word Error Rate* (WWER) for its evaluation metric. WWER weights some words or word types differently than it does others. In IRIS, we choose to weight each personal name piece as five points and each locative name piece as two points. All other words are treated as one point. Any genealogical information that is written in the header is given half as much weight as just described, but the header is treated as being attached to each row of content.

4.2 Flexible Evaluation Systems

In evaluating system performance, IRIS seeks to be as generous as possible. Consequently, it has built an evaluation tool which attempts to maximize a system's score despite potential row or word segmentation issues ; and which accounts for the fact that the human transcripts themselves may have errors or may not represent the actual apparent data due to constraints of the guidelines which were created for purposes other than IRIS. We here describe this flexible evaluation system.

4.2.1 Flexible Handling of Segmentation Issues

Within a properly-identified cell, the HR system may hypothesize "George E" when the truth is "George." Moreover, and perhaps more catastrophic, since systems must find their own bounding boxes, it is possible to skip or partially transcribe rows and columns. For example, a system could fail to transcribe the first row of a 50-row census image and then perfectly transcribe the remaining rows. Without scoring flexibility, this might appear as 50 deletions and 49 insertions. Yet a human marking the system might say it only had 2% error.

To account for segmentation errors *within a given cell*, IRIS uses minimum edit distance to attempt to provide as much value to the cell transcript as possible. In the case of "George E." versus "George," it will credit the system with having found "George" even though the system postulated additional information. This type of flexibility is common in recognition evaluation systems – we wish mainly to focus on the fact that the flexibility happens at the cell level. Using this, for a given hypothesis row *h* and a reference row *r*, the tool should be able to determine an optimally-scoring alignment between *h* and *r*.

Since the scorer can align any hypothesis row to any reference row, one can use the scorer to compute a minimum edit distance across all *rows*. So in the case above where the HR system failed to transcribe a first row, the flexible scorer would also report a "2%" error. That said, we realized that if an HR system fails to recognize its appropriate *column* (eg., treating a person's age as gender), such an error would be harder for downstream systems to leverage. Therefore, the scorer does not accommodate column segmentation failures.

4.2.2 Handling Gold Standard Idiosyncrasies

The gold standard has issues of its own. These issues should not yield penalties for the recognition systems. Hence, the scorer was prepared to handle these idiosyncrasies. In particular, these issues involve (a) forced transcription choices, (b) handling of dittos, (c) variations in word segmentation, and (d) word choice.

As was mentioned earlier, a census may have birth origin recorded as "PA." Yet the transcriber may have recorded the origin as "Pennsylvania." This is due in part to the fact that when volunteers transcribe documents, they are told in some cases to transcribe what they see unless it is an abbreviation and they can determine what the abbreviation means. So a transcriber might appropriate select Pennsylvania for this case. Furthermore, the transcription tool uses authority tables to highlight potential transcription concerns which may intimidate annotators into transcribing Y when the correct transcript would have been X but was previously unobserved in the tables. Similarly, there are interpretational issues that occur. "Hernandez" may appear to be "Hernandes;" or "McDonald" with no space may appear in the image as "Mc Donald" – now with a space. Since these kinds of differences do not affect usability, the scorer was fitted with lists of acceptable variations of these kinds.

Lastly, transcribers are also taught that if they see ditto marks or the equivalent thereof, they should transcribe the information that was intended. So "Smith" may occur on one line, and "*do*", meaning ditto, on the next line … which would be transcribed again as "Smith." To handle this, our scorer was instructed that if the HR system gave no result where a ditto occurred, it would not be penalized; but if it said "Smith" where a ditto of "Smith" should occur, it will get credited with a correct value.

5 Evaluation Results

5.1 Current Best Automatic Performance

As was mentioned earlier, several major image research organizations have been able to produce results on the IRIS collection. Specifically, these have been BBN and A2ia, which have been key players in MADCAT and OpenHaRT. The best-performing per-collection results from these participants can be illustrative of the difficulty of the various tasks and of the performance metrics other researchers would need to achieve in order to advance the state of the art. Table 2 shows the best per-collection WWER scores to date, as well as best and worst case results on any particular collection.

Collection	Average Per-Record WWER	Best Case WWER	Worst Case WWER	Std Dev Per-Record WWER
1930 US Census	**19.6%**	2.73%	98.4%	12.8%
1930 Mexico Census	**47.4%**	5.59%	374%	30.9%
Arkansas Marriages	**29.4%**	0.00%	103%	18.4%
French Parish	**92.4%**	22.0%	198%	12.1%

Table 2: Best-performing System WWER per Collection

5.2 Human Levels of Performance

The primary goal of IRIS was to see how close HR is to replacing some/all human transcription of historical documents. So it is relevant to know how close these scores are to human levels of performance.

Collection	Average Per-Record WWER (Human)	Automatic / Human Ratio of Average Per-Record WWER
1930 US Census	**7.9%**	2.48

Table 3: Human WWER on US Collection

As was previously stated, each IRIS transcript was a product of three separate volunteers: two volunteers, "A" and "B," who transcribed the document independently; and an arbitrator (ARB). The gold standards for each collection except the US census were derived from ARB transcripts. However, as was mentioned, the US 1930 Census was re-transcribed to 99.7% accuracy. This gold standard could be therefore used to evaluate human

accuracy as well. Table 3 shows the human WWER for side-B transcripts. Note that humans are only about 2.48 times better than automation which suggests that automation is not far from reaching human accuracies.

5.3. Per-Category Results

Further research is clearly needed to move error rates down to human levels. Even so, analysis reveals that HR might be ready for production usage with some *fields*. That is, a system could be useful for transcribing a particular field if it either has high *accuracy* (1-WER) or, if whenever it makes a hypothesis, it is typically correct (i.e., it has high *precision*). Table 4 shows IRIS participant's current best accuracies and precisions for each field in the 1930 US Census. Table 4 shows that recognition on small vocabulary fields performs well.

US 1930 FIELD	ACCURACY	PRECISION
census_district (H)	0.362	0.373
census_county (H)	0.649	0.741
sheet_number (H)	0.724	0.742
sheet_ltr (H)	**0.976**	**0.992**
household_id	0.747	0.825
pr_name_full	0.813	0.840
pr_relationship	**0.910**	**0.940**
pr_sex	**0.943**	**0.961**
pr_race_or_color	**0.946**	**0.969**
pr_age	0.840	0.857
marital_status	**0.939**	**0.957**
pr_birthplace	0.757	0.864
pr_fthr_birthplace	0.771	0.874
pr_mthr_birthplace	0.776	0.877

Table 4: Field performance for Best US Census System

This same trend continues in small vocabulary fields of other collections, as is shown in Table 5A (Mexico collection), Table 5B (Arkansas collection), and Table 5C (Parish records). Perhaps automation could be applied *today* for these particular fields. Nonetheless, we hope that making this collection available to the research community will lead to improvements in the larger-vocabulary fields.

MEXICO FIELD	ACCURACY	PRECISION
Marital status	0.893	0.989
Principal's Age	0.797	0.919
Relation to Head	0.821	0.999
Gender	0.843	0.983

Table 5A:Best of Mexico Census Small Vocabulary Fields

ARKANSAS FIELD	ACCURACY	PRECISION
Event Type	0.898	0.950

Table 5B:Best of Arkansas Small Vocabulary Fields

ARKANSAS FIELD	ACCURACY	PRECISION
Event Month	0.232	0.754
MultRecord Type	0.428	0.730
Gender	0.296	0.740

Table 5C:Best of Parishes' Small Vocabulary Fields

6 Error Analyses

The evaluation results only provide an overall view about systems. It is beneficial to likewise drill down and determine what kinds of errors are being made. We will here to focus on major sources of error.

The systems that have been evaluated on IRIS have generally had success with transcription of the US and Mexico 1930 censuses. Yet key errors have occurred when the original scripts were too light (as if written in pencil), where there were occlusions, and when the number of people listed on a given census page was significantly fewer than the 50 that were possible.

6.1 Faint Images

If we sort the images by identifiers, they end up being largely grouped by the particular state in which the census was conducted. Figure 5 shows WWER of the US 1930 Census according to this method of sorting, where 1.0 on the Y-axis indicates 100% weighted word error.

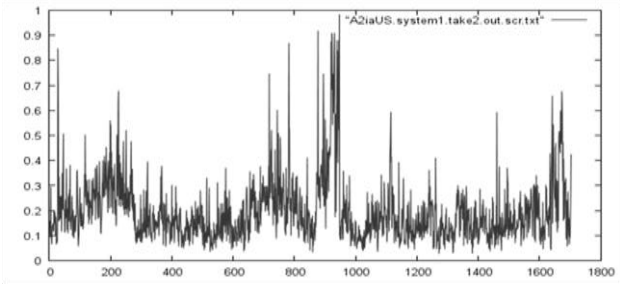

Figure 5: WWER on US Census in Index-sorted Order

In this image, it is clear that there is a region of documents where system performance regularly approaches 100% WWER. If we look deeper at these particular documents, we note that all of them came from a jurisdiction in Vermont. Figure 6 shows an instance of an image from this portion of Vermont.

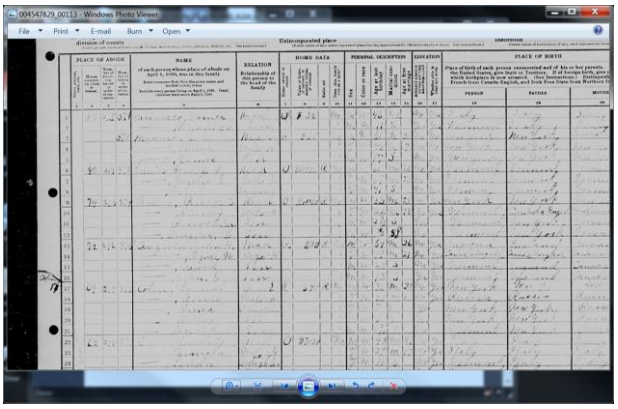

Figure 6: Instance of Low-Readability Images

When we first noted these anomalies, we initially thought that the images were completely blank. On closer inspection, we realized that the images were readable by humans, but it is no surprise that they would be difficult for systems to interpret. This phenomenon led to almost a 1% absolute reduction in WWER for the best system.

6.2 Occlusions and Empty Cells

Again, in looking at an index-sorted order of images from the Mexico Census, we observed the same kinds of error regions as we had seen with the US data. In these regions, WWER was sometimes in excess of 300%: catastrophic failures. In looking at image types that were causing these problems, some were due to occlusions in the image (such as Figure 7) which made the system believe there was more text on the pages than needed to be transcribed.

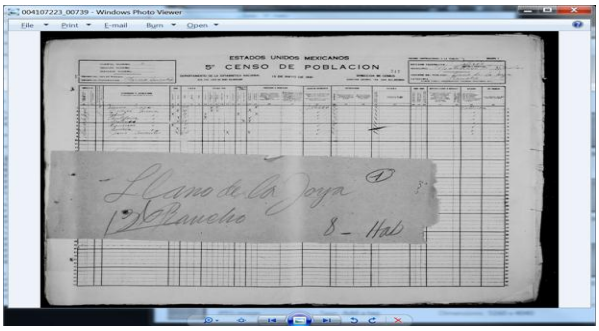

Figure 7: Occluded Images Result in High WWERs

The second similar phenomenon is shown in Figure 8. In cases like Figure 8, where there are only two transcribable lines of data, some of the systems reported 50 results which, then, resulted in significant insertion penalties.

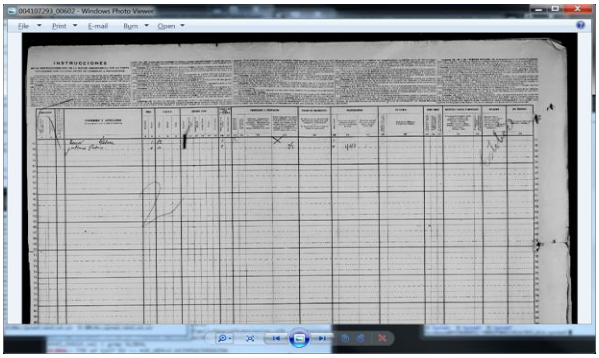

Figure 8: Sparse Tables Yield High Insertion Rates

These types of errors also resulted in close to a full percentage point of WWER. Thus elimination of these major errors would bring systems that much closer to human levels of accuracy.

7 Conclusions

This IRIS evaluation demonstrates that handwriting recognition systems could presently be applied to genealogically-relevant small vocabulary tasks. With some perhaps limited additional research, full transcription of genealogical records seems to be a near-term reality. We are eager to release the IRIS collection and tools to the research community to help bridge the final gap between human transcription and automation.

8 References

"Arkansas Marriages, 1837-1944" (accessed 2014). Index: FamilySearch (http://FamilySearch.org), based on data collected by Genealogical Society of Utah, Salt Lake.

DARPA (2008). Multilingual Automatic Document Classification, Analysis and Translation (MADCAT). http://www.darpa.mil/Our_Work/I2O/Programs/Multilingual_Automatic_Document_Classification,_Analysis_ and_Translation_%28MADCAT%29.aspx

España, S., Castro, M.J., Hidalgo, J.L. (2004). The SPARTACUS-Database: a Spanish Sentence Database for Offline Handwriting Recognition. *Proc. of LREC 2004*, pp. 227-230.

FamilySearch (2014). FamilySearch Indexing.

"France, Diocèse de Coutances et d'Avranche, registres paroissiaux, 1533-1906" (accessed 2014). Index and images: FamilySearch (http://FamilySearch.org).

Goldburg, H.E. (1914). *Controller*. US Patent 1,117,184.

Hansel, C W. (1939) *Multiplex facsimile printer system*. U.S. Patent 2,143,875.

Lengel, E.G./editor (2008). *The Papers of George Washington, Digital Edition,* Univ. of Virginia Press.

Liu,C., Yin F., Wang, D., Wang, Q (2011) CASIA Online and Offline Chinese Handwriting Databases. Proc. of *2011 ICDAR*, pp. 37-41

Marti, U.-V., Bunke, H. (2002) The IAM-database: an English sentence database for offline handwriting recognition. *Int. J. Document Analysis and Recognition*, 5(1): 39-46, 2002.

"México, Censo Nacional, 1930" (accessed 2014). Index and Images: FamilySearch (http://FamilySearch.org), citing Instituto Nacional de Estadísticam Geografía e Informática. Achivo General de la Nación, Distrito Federal.

NIST (2010). *NIST 2010 Open Handwriting Recognition and Translation Evaluation Plan*,www.nist.gov/itl/mig/upload/OpenHaRT2010_EvalPlan2010_EvalPlan_v2-8.pdf

NIST (2013). *OpenHaRT 2013 Evaluation Results*. ftp://jaguar.ncsl.nist.gov/outgoing/HART13_REPORTS_20131204/index.html

Song, Z., Ismael, S., Grimes, S., Doermann, D., Strassel, S. (2012). Linguistic Resources for Handwriting Recognition and Translation Evaluation. In *Proceedings of LREC 2012*, pp. 3951-3955.

Strassel, S.M. (2009). Linguistic Resources for Arabic Handwriting Recognition. Second International Conference on Arabic Language Resources and Tools.

Su, T.H., Zhang, T.W., Guan, D.J. (2007) Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text, *Int. J. Document Analysis and Recognition*, 10(1): 27-38, 2007.

"United States Census, 1930" (accessed 2014) Index and images: FamilySearch (http://FamilySearch.org), citing National Archives and Records Administration Publication T626, 2002. Washington, D.C.

Viard-Gaurdin, C., Lallican, P.,M., Knerr, S., Binter, P. (1999). The IRESTE On/Off (IRONOFF) dual handwriting database, Proceedings of 5^{th} *ICDAR* , pp. 455-458.

Wikipedia:FamilySearch (accessed 2014). en.wikipedia.org/wiki/familysearch