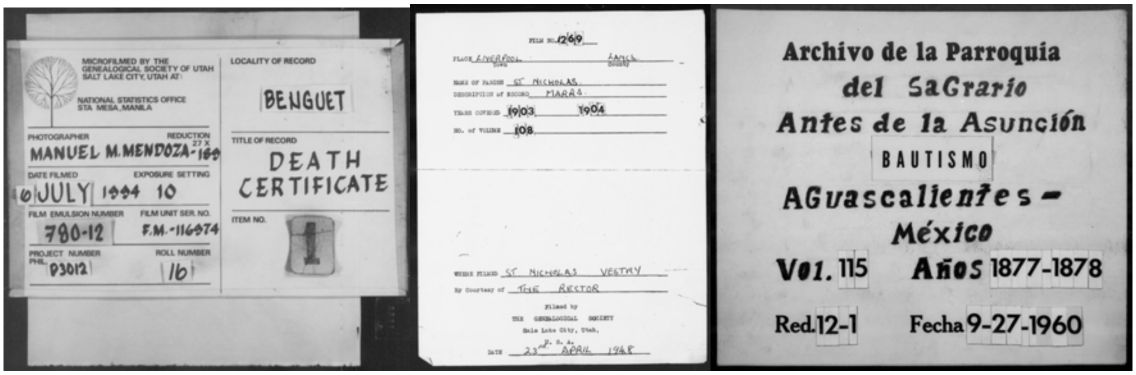
A System to Automatically Index Genealogical Microfilm Titleboards

Samuel James Pinson, Mark Pinson and William Barrett
Department of Computer Science
Brigham Young University

Introduction

Millions of rolls of microfilm contain valuable genealogical information, yet remain largely inaccessible. A titleboard (see Figure 1) contains semi-structured metadata about the genealogical records that follow the titleboard on the roll of microfilm. This metadata may include the geographical origin of the records (i.e. city and country), the record type (i.e. birth records, marriage records), and relevant dates.

We propose a system to automatically segment, extract, index and search digitized titleboards. Titleboards, the input to our system, pass through three modules (preprocessing, text recognition, and indexing) to produce an index over a digitized microfilm library. The index is accessed through a graphical user interface via boolean and regular expressions. The research we present here focuses on a necessary step in the indexing process: method identification. Method identification is the process of determining whether text is machine print or handwriting.



MICROFILMED BY THE
GENEALOGICAL SOCIETY OF UTAH
SALT LAKE CITY, UTAH AT:
NATIONAL STATISTICS OFFICE
STA. MESA, MANILA

PHOTOGRAPHER MANUEL M. MENDOZA — REDUCTION 27 X - 16
DATE FILMED 6 JULY 1994 — EXPOSURE SETTING 10
FILM EMULSION NUMBER 780-12 — FILM UNIT SER. NO. F.M.-116974
PROJECT NUMBER PHIL. P3012 — ROLL NUMBER 16
LOCALITY OF RECORD BENGUET
TITLE OF RECORD DEATH CERTIFICATE
ITEM NO. 1

FILM NO. 1269
PLACE LIVERPOOL (TOWN) LANCS. (COUNTY)
NAME OF PARISH ST NICHOLAS
DESCRIPTION of RECORD MARRS.
YEARS COVERED 1903 1904
NO. of VOLUME 108
WHERE FILMED ST NICHOLAS VESTRY
By Courtesy of THE RECTOR
Filmed by
THE GENEALOGICAL SOCIETY
Salt Lake City, Utah,
U. S. A.
DATE 23rd APRIL 1948

Archivo de la Parroquia
del SaGrario
Antes de la Asunción
BAUTISMO
AGuascalientes -
México
Vol. 115 Años 1877-1878
Red. 12-1 Fecha 9-27-1960

Figure 1: Typical microfilm titleboards. Titleboards preface a collection of frames on a roll of microfilm and describe the type of information to be found in the succeeding frames. Segmentation and extraction of the content of the titleboards provides native information vital to the creation of a meaningful index for the frames that will follow, greatly increasing user access to the information contained in those frames.

Preprocessing

The preprocessing stage performs a sequence of steps on grayscale images of microfilm titleboards. This includes noise removal and locally adaptive thresholding [1]. Connected components, or groups of adjacent black pixels, are extracted from the binarized images.

Method Identification

Turk and Pentland published a face recognition technique called *Eigenfaces* [2]. We may consider an NxN image as a point in an N^2-dimensional vector space. Eigenfaces relies on discovering *face space*, the subspace that best represents the distribution of images of faces in the N^2-dimensional space. Images of known faces as well as images of faces to be recognized are projected into face space. The minimum Euclidean distance from an unknown face to the known faces determines how the unknown face will be classified. If the minimum distance is above some threshold, then the face is rejected, i.e. not recognized.

The basis of face space consists of the most significant eigenvectors of the covariance matrix of a training set of face images. The covariance matrix is given by

$$C = \frac{1}{M}\sum_{i=1}^{M}(\Gamma_i - \Psi)(\Gamma_i - \Psi)^T \ ,$$

where Γ_i is the ith training image and Ψ is the average of all the training images. The projection of a face, Γ, into face space is a vector of weights, one for each dimension of face space. The kth component of the projection is $\omega_k = u_k^T(\Gamma - \Psi)$, where u_k is the kth most significant eigenvector of the covariance matrix C.

Muller and Herbst [3] employ this idea in character recognition. Images of faces are replaced with images of characters. We build on their work and apply the principle to connected component level method identification. That is, for each connected component in an image, we determine whether the connected component is machine print or handwriting.

We accomplish this by determining the face space for a large set of machine print characters. Representative machine print characters are then projected into this space. We determine a local distance threshold for each representative machine print character based on a user-supplied global requirement for machine print precision and the radial density (See Figure 2) of machine print and handwriting surrounding the connected component. This algorithm is outlined below.

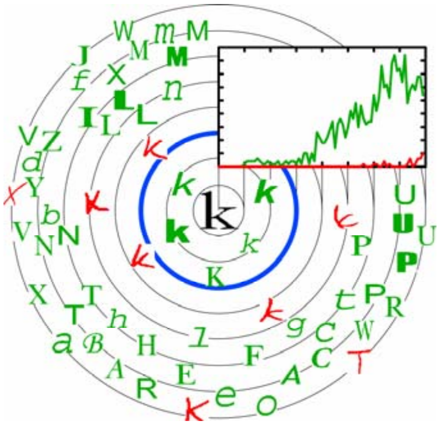


Figure 2. Radial density

θ^{global} : user-supplied global requirement for machine print precision.

θ_i^{local} : local distance threshold for the ith machine print representative.

$r^{mp}(d)$: number of machine print connected components within distance d of the ith machine print representative.

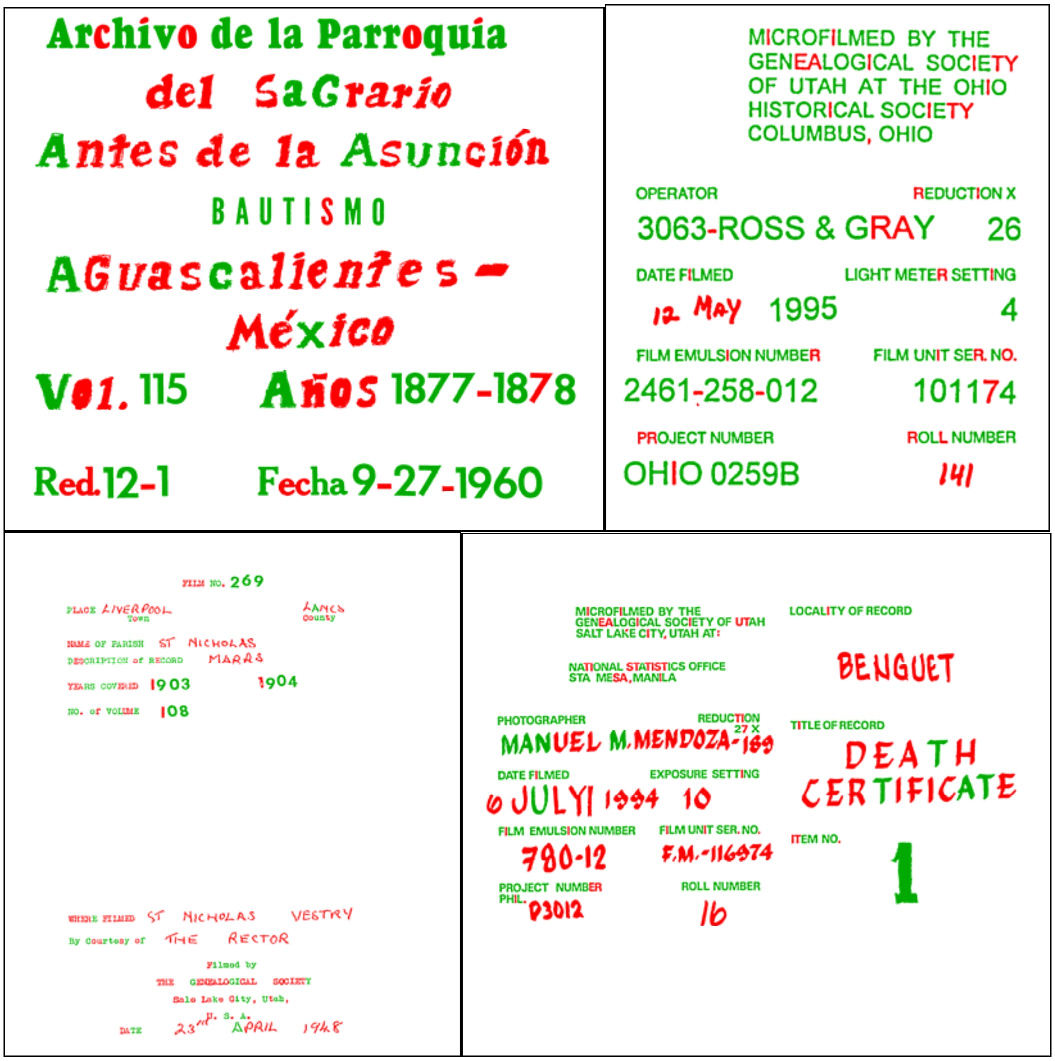
$r^{hw}(d)$: number of handwriting connected components within distance d of the ith machine print representative.

```
1. $\theta_i^{local} = 0$
2. while ( $r_i^{mp}(\theta_i^{local}) / [r_i^{mp}(\theta_i^{local}) + r_i^{hw}(\theta_i^{local})] \geq \theta^{global}$ )
        Increment $\theta_i^{local}$
3. if ( $\theta_i^{local} > 0$)
        Decrement $\theta_i^{local}$ .
```

Thus, for each representative machine print connected component, the local distance threshold grows from zero to just before the global requirement for machine print precision first fails to be satisfied.

Connected components that are within the local threshold of their nearest machine print representative in face space are classified as machine print. Any connected components lying beyond the local threshold are deemed handwriting. For example, in the annular diagram handwriting connected components begin to appear at the third ring. Increasing the local threshold further will drop the machine print precision below the global requirement, so the local threshold is fixed at the third ring (shown in blue).

Preliminary results on microfilm titleboards are depicted in Figure 3 below.



Archivo de la Parroquia
del SaGrario
Antes de la Asunción
BAUTISMO
AGuascalientes -
México
Vol. 115 Años 1877-1878
Red. 12-1 Fecha 9-27-1960

MICROFILMED BY THE
GENEALOGICAL SOCIETY
OF UTAH AT THE OHIO
HISTORICAL SOCIETY
COLUMBUS, OHIO
OPERATOR 3063-ROSS & GRAY
REDUCTION X 26
DATE FILMED 12 May 1995
LIGHT METER SETTING 4
FILM EMULSION NUMBER 2461-258-012
FILM UNIT SER. NO. 101174
PROJECT NUMBER OHIO 0259B
ROLL NUMBER 141

FILM NO. 269
PLACE LIVERPOOL Town LANCS County
NAME OF PARISH ST NICHOLAS
DESCRIPTION of RECORD MARRS
YEARS COVERED 1903 1904
NO. of VOLUME 108
WHERE FILMED ST NICHOLAS VESTRY
By Courtesy of THE RECTOR
Filmed by
THE GENEALOGICAL SOCIETY
Salt Lake City, Utah,
U. S. A.
DATE 23rd APRIL 1948

MICROFILMED BY THE
GENEALOGICAL SOCIETY OF UTAH
SALT LAKE CITY, UTAH AT:
NATIONAL STATISTICS OFFICE
STA MESA, MANILA
LOCALITY OF RECORD BENGUET
PHOTOGRAPHER MANUEL M. MENDOZA-169
REDUCTION 27 X
DATE FILMED 6 JULY 1994
EXPOSURE SETTING 10
FILM EMULSION NUMBER 780-12
FILM UNIT SER. NO. F.M.-116974
PROJECT NUMBER PHIL. 03012
ROLL NUMBER 16
TITLE OF RECORD DEATH CERTIFICATE
ITEM NO. 1

Figure 2: Preliminary method identification results on microfilm titleboards. The connected components in these titleboards have been classified by our system. Green signifies machine print, while red indicates handwriting.

The confusion matrix over these four titleboards is:

	Predicted handwriting	Predicted machine print
Actually handwriting	**88.9%**	11.1%
Actually machine print	16.6%	**83.4%**

Titleboards are a particularly noisy type of document to process. Thus, the results presented here are below the system's typical performance. We also note that 35.2% of our handwriting false positives were caused by touching machine print characters. Thus, one way to improve our performance is to split touching characters.

The upper left image in Figure 2 reveals a difficulty in method application on titleboards. It is evident that the writer expended great effort to make his or her handwriting actually look like machine print. Despite this, our algorithm is able to correctly distinguish in most cases.

Document Segmentation

The Docstrum [4] layout analysis algorithm groups the connected components into text lines and words. These results of method identification presented above are especially encouraging since we are making our classification at the connected component level. Robust rules may be developed to make classifications at the word, text line, or block level.

Character Identification

Following method identification, the text is passed to the appropriate (either machine print or handwriting) character recognition engine.

Index Construction

Recognition errors are inevitable due to poor image acquisition, image degradation, segmentation errors, etc. This must be considered during index construction. We must decide whether to directly index possibly incorrect recognition results or attempt to correct the recognition results. Lexical context may serve to correct recognition errors, requiring, for example, that indexed terms must be present in a predefined dictionary.

Information retrieval systems are typically evaluated by two measures: precision, and recall. Precision is defined as the percentage of the retrieved results that are actually relevant. Recall is defined as the percentage of the relevant results that were actually retrieved. Ideally, we would like 100% precision and 100% recall. In practice, we have to choose between the two. The nature of our problem, genealogical research, demands high recall at the expense of some precision.

Thus, we build an index based directly on the (possibly incorrect) recognition results, rather than trying to force the indexed terms to be dictionary terms.

Querying

Queries take the form of Boolean or regular expressions in our system. However, these queries are expanded using approximate string matching to increase the recall. Approximate string matching seeks to find the closest match for a word based on a set of editing operations. Each operation, such as "insert a character" or "delete a character", has an associated cost. The query expansion is user controlled via an adjustable cost threshold. For example, the index might contain the term "buriais", stemming from misrecognizing "burials". The query "burials" is expanded to include all keywords within some edit distance from "burials".

Future Work

The focus of this paper has been our research in connected component level method identification. We have implemented a prototype of the framework proposed, but much work remains incomplete before a robust system will be available. This work may also be extending in many ways. Adding metadata to index terms such as script, language, and meaning would enhance searching capabilities. For example, we could then search for titleboards with German city names.

Our system makes classifications at the connected component level. Exploitation of spatial, stylistic, font, and linguistic context promises to increase robustness and accuracy. For example, we may safely assume that the connected components within a word are either all handwriting or all machine print. Confidence-weighted classifications of the constituent connected components ought to be harmonized with this assumption. Similar reasoning holds for the style and font of connected components within a word. Likewise, linguistic context in the form of lexicons and n-gram frequencies may also be brought to bear if character recognition is performed.

Finally, we note that the system could be improved by specializing the set of representative machine print connected components. Currently, the representative set consists of 5957 machine print connected components in a wide variety of fonts and styles. This is necessary for success over a broad range of text. However, limiting the size of this set will increase the speed of the algorithm and simultaneously decrease the rate of handwriting false positives. In particular, for indexing microfilm titleboards the set of representative fonts could be limited to those fonts that are present in a random, but representative, sample of titleboards.

Conclusion

We have proposed a system for automatically indexing genealogical microfilm titleboards. We have implemented a proof-of-concept prototype, and the system remains a work in progress. However, the proposed system promises to powerfully improve genealogical research by creating a searchable index of microfilm titleboards.

References

[1] **Evaluation of binarization methods for document images**, Trier, O.D.; Taxt, T. Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol.17, Iss.3, Mar 1995, pp. 312-315

[2] **Face recognition using eigenfaces**, Turk, M.A.; Pentland, A.P. Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on, Jun 1991, pp. 586-591

[3] **The use of eigenpictures for optical character recognition**, Muller, N.; Herbst, B. Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on, Vol.2, Aug 1998, pp.1124-1126

[4] **The document spectrum for page layout analysis**, O'Gorman, L. Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol.15, Iss.11, Nov 1993, pp.1162-1173