Line Detection in Binary Document Scans: A Case Study with the International Tracing Service Archives

Benjamin Charles Germain Lee

2017 Digital Humanities Associate Fellow
Levine Institute for Holocaust Education &
Jack, Joseph, & Morton Mandel Center for
Advanced Holocaust Studies
United States Holocaust Memorial Museum

Visiting Fellow
Department of History
Harvard University

Source: 20th Anniversary of the United States Holocaust Memorial Museum, “International Tracing Service,”
The International Tracing Service (ITS)

- Established “to help reunite families separated during [World War II] and to trace missing family members”\(^1\)
- Invaluable resource for Holocaust survivors and their families, as well as Holocaust researchers
- Contains document subcollections ranging from concentration camp records to correspondences attempting to identify unaccompanied children\(^2\)
- Digitization effort started in 2007\(^3\)
- Over 190 million images in the digitized archive\(^3\)

Source: 20\(^{th}\) Anniversary of the United States Holocaust Memorial Museum, “International Tracing Service,”
What is the Central Name Index?

- Tracing service -> archival material indexed by name
- **47+ million cards** in the Central Name Index (CNI)
- Significant for two reasons:
  - Constitutes the central finding aid for the collection
  - Contains certain document types of historical interest, such as Death Certificates issued by ITS, that can only be found in the CNI

The Physical Collection of Central Name Index cards

Barriers to Research

- Only names are associated with the images of the CNI cards in the ITS digitized archive, meaning that the user cannot search the archive by fields such as nationality, concentration camp, age, etc.
- Difficult to extract additional metadata from the images because the scans are low resolution binary TIFs with handwriting prevalent throughout.
- No way to search for specific card types of historical interest within the CNI (would require a manual search, which is entirely intractable).

CNI Card Types

• Can be classified by card type based on the form structure of the card (in the spirit of document layout analysis)
• Form structure oftentimes indicates card function (e.g., what type of document the card references)
• Inspiration for this: Card Guide for the Holocaust Survivors and Victims Resource Center, United States Holocaust Memorial Museum, written by Jo-Ellyn Decker, Sara-joelle Clark, and Laura Ivanov

Sterbeurkunde (Death Certificate)

(info redacted for privacy considerations)

Sonderstandesamt Arosen (Special Civil Registry Office Arosen) stamp indicating this is an official ITS issued death certificate.

Single Line Cards

These cards often act as reference cards (see pages 66-68). However, some single line cards are also inquiry cards (see page 17). Others provide "Internal Hints" (see page 38).

Multi-Line Cards

Type A

Type B

Type C

Multi-line cards are often reference cards (see pages 66-68). However, some older inquiry cards (see page 17) and original cards (see page 43) also utilize multi-line formats.

T-Line Cards

T-Line cards reference an original document. An individual's information will appear at the top of the card.

Complex Line Cards

No-Line Cards

Complex line cards are often reference cards (see pages 66-68), such as Suchstelle cards (see page 109), but can also be reproductions of original documents, including Gestapo cards or ITS-created envelopes (see page 43).

No-line cards may act as inquiry cards (see page 17), reference cards (see pages 66-68), or present internal hints (see page 38).

Question:

Can we sort the 47+ million CNI cards based on their line structure and form type in order to isolate card types of historical interest?
Line Detection for Classification

- First method
- Detailed in short paper

Source: 20th Anniversary of the United States Holocaust Memorial Museum, “International Tracing Service,”
Source: 20th Anniversary of the United States Holocaust Memorial Museum, “International Tracing Service,”
New Method: Template Matching + Random Forest

• Exploit form structures of different card types
• Isolate template fields such as “Name” for each card type
• Cross-correlate templates with cards
• Retain n x 3 matrix of n (x, y, max) tuples (consisting of the normalized x & y positions of maximum of cross-correlation, as well as the maximum value itself)
• Train Random Forest on card matrices with labels
• Run cross-correlation on cards without labels, generate matrices, use Random Forest to predict labels

Test Case: 5 card types, 4580 cards

- Classified >5,000 cards by hand
- Selected 5 prevalent card types to begin classification tests:
  - AEF (101 cards)
  - Multiline (865 cards)
  - No line (1203 cards)
  - T (2300 cards)
  - T with boxes (111 cards)
- 77 templates in total
- Randomly separate into training set and test set

Source: 20th Anniversary of the United States Holocaust Memorial Museum, “International Tracing Service,”
Failure Mode: T / multiline boundary:

Name:

geb.

am:
in:

Nat.:

Az.:

Oräner Nr.

Seite:

(information redacted for privacy considerations)
Next Steps

• Scale algorithm to run on all card types, all 47+ million card scans in digitized CNI
• Improve runtime from 1 card/second
• CPU / GPU
• Isolate Death Certificate cards for the first time
  – unique centralized death certificate record for people in camps
• Train neural network to read information off of cards
• Novel historical research with death certificate cards!

Source: 20th Anniversary of the United States Holocaust Memorial Museum, “International Tracing Service,”
Dem Sonderstandesamt vorgelegt

Name: Komieczny, Marian

geboren am 11. 7. 1922 in Gorzewiec

Nat. Polen

gestorben am 8. 4. 1943, 9.00 in Mauthausen

Todesursache a. d. Flucht erschossen

beerigt am ......................................................... in .................................................................

Sterbeurkunde Nr. 826 Abt M

Datum: 26. 2. 52

ausgestellt aufgrund folgender Dokumente

Bitte wenden

CNI card of Marion Komieczny, 0.1/28004164/ITS Digital Archive, USHMM
Acknowledgments

• International Tracing Service

• United States Holocaust Memorial Museum
  – Michael Haley Goldman, Robert Ehrenreich, Michael Levy, Jo-Ellyn Decker, Sara-joelle Clark, Laura Ivanov, Elizabeth Anthony

• Harvard University Department of History
  – Gabriel Pizzorno

Works Cited

1) The Holocaust Survivors and Victims Resource Center at the United States Holocaust Memorial Museum, “ITS Frequently Asked Questions”  

2) Charles-Claude Biederman, 60 Years of History and Benefit of the Personal Documentary Material about the Former Civilian Persecutees of the National Socialist Regime Preserved in Bad Arolsen

3) Senate Resolution 142 (2007): “A resolution observing Yom Hashoah, Holocaust Memorial Day, and calling on the remaining member countries of the International Commission of the International Tracing Service to ratify the May 2006 amendments to the 1955 Bonn Accords immediately to allow open access to the Bad Arolsen archives.”


5) CNI card of Marion Komieczny, 0.1/28004164/ITS Digital Archive, USHMM

6) CNI card of Zbigniew Pracownik, 0.1/32699964/ITS Digital Archive, USHMM

7) 20th Anniversary of the United States Holocaust Memorial Museum, “International Tracing Service,”  

8) Source: United States Holocaust Memorial Museum, “View of the Central Name Index of the International Tracing Service in Bad Arolsen,”  
https://collections.ushmm.org/search/catalog/pa1165346

Source: 20th Anniversary of the United States Holocaust Memorial Museum, “International Tracing Service,”  