Machine learning, template matching, and the International Tracing Service digital archive: Automating the retrieval of death certificate reference cards from 40 million document scans

Benjamin Charles Germain Lee
Levine Institute for Holocaust Education and Jack, Joseph and Morton Mandel Center for Advanced Holocaust Studies, United States Holocaust Memorial Museum, USA and
Visiting Fellow, Department of History, Harvard University Cambridge, USA

Abstract

Scattered throughout the International Tracing Service (ITS) digital archive, one of the largest and most heterogeneous collections of Holocaust-related material, are hundreds of thousands of reference cards to official death certificates recording a fraction of individuals who perished within concentration camps. These cards represent the most comprehensive collection of digital material pertaining to these death certificates issued by Sonderstandesamt Arolsen, a German civil registry office. However, the reference cards can only be found dispersed throughout the Central Name Index (CNI), ITS's 46+ million-card finding aid that is indexed only by name. Consequently, aggregating the death certificate reference cards for research requires an intractable manual search. I adopt template matching and machine learning to automate the retrieval of these cards from the ITS digital archive. I demonstrate the efficacy of my method on a test set of 22,117 hand-classified cards, reporting 100% precision and 100% recall. Running this algorithm on 39,967,358 scans of cards from the CNI, I identify 312,183 death certificate reference cards in 13.75 days of elapsed real runtime on a personal computer with only a single, $600 Intel processor. Finally, I demonstrate that this approach can be generalized to many different card types within the CNI, showing great promise for application to other archives.

Correspondence:
Benjamin Charles Germain Lee, CSE 424, Paul G. Allen Center for Computer Science & Engineering, University of Washington, 185 E Stevens Way NE, Seattle, WA 98185, USA.
E-mail:
bcgl@cs.washington.edu

doi:10.1093/llc/fqy063 Advance Access published on 9 November 2018

1 Introduction

A prevalent problem in archival research is isolating documents of interest in the absence of an adequate finding aid. In the case of digital archives, this problem manifests most prominently in trying to retrieve archival material without sufficient electronic metadata. This problem can be devastatingly restrictive: if a researcher is unable to query an archive according to desired fields, the researcher is forced to search through the material manually. However, in the case of modern digital archives, the number of scanned documents can reach the hundreds of millions, meaning that a manual, brute-force search of even just a small fraction of the archive is infeasible. An ideal solution to this problem would be to automate the extraction of metadata from scanned documents. Indeed, this has become one of the most preeminent challenges of modern archival science, and a diverse range of solutions has been explored, including optical character recognition (OCR), handwriting recognition, and document layout analysis. In this article, I focus on one subproblem: isolating documents with the same layout structure from a larger, heterogeneous corpus of scanned documents. The specific case study that I address is the International Tracing Service (ITS) digital archive, one of the world's most voluminous repositories of documents related to the Holocaust, where documents of interest such as the reference cards to Sonderstandesamt Arolsen death certificates are strewn throughout with no method of retrieval other than a manual search.

1.1 The ITS archive

ITS was established with the goal 'to help reunite families separated during [World War II] and to trace missing family members' (ITS FAQ, 2007). It had its genesis in 1943, when the British Red Cross Bureau for International Affairs was converted into a tracing service (Brown-Fleming, 2016, p. 3). This initial tracing service would survive many bureaucratic reorganizations, eventually being moved to Bad Arolsen, Germany, in 1946, renamed the International Tracing Service in 1948, and subsumed by the International Committee of the Red Cross in 1955, at which time an International Commission became its governing body (Brown-Fleming, 2016, pp. 3–4). Currently, ITS still functions as a tracing service in Bad Arolsen for those seeking information on specific victims of the Holocaust, and its governing International Commission has grown to eleven member countries (Agreement on the International Tracing Service, 2011, p. 3). In the words of Paul Shapiro, former Director of the United States Holocaust Memorial Museum's (USHMM) Center for Advanced Holocaust Studies, ITS's holdings:

> [relate] to the fates of millions of people, Jews and members of virtually every other nationality as well, who were victimized by the Nazis: millions of concentration camp documents; transport and deportation lists; Gestapo arrest warrants; prison records; forced and slave labor documentation ... displaced persons (DP) files; and millions of inquiries from around the world from survivors and their families, all hoping at first to find someone still alive, and later simply needing to better understand what had happened to loved ones who had been murdered. (Brown-Fleming, 2016, p. x)

Indeed, ITS has accrued an extremely heterogeneous collection of documents over the past seven decades. In addition, many of these documents are unique, meaning that ITS has a uniquely important role in Holocaust documentation (McDonald, 2007, p. 1362).

Even though ITS continued to serve as a tracing service in the new millennium, the archive remained closed to survivors, victims, their families, historians, and other researchers, amassing over 400,000 unprocessed requests by 2006 (Belkin, 2007, p. 1). This severely restricted Holocaust research, obfuscated and slowed the tracing process, and embargoed an enormous corpus of incontrovertible documents that could be used to combat Holocaust denial (Brown-Fleming, 2016, p. xi). Consequently, governmental agencies and institutions such as the USHMM and the United States Senate placed increasing pressure on the eleven country International Commission to ratify an amendment to open the archive, including passing Senate Resolution 142 of the 110th Congress calling

for other member states to ratify the amendment (Senate Resolution 142 – 110th Congress, 2007). The final logistics of the amendment were completed in 2007, and copies of the digital archive were given to one member institution in each of the eleven countries in the International Commission: for example, the USHMM and Yad Vashem became the holders of the ITS digital archive in the USA and Israel, respectively. As a result, survivors, family members, and researchers can now access the ITS digital archive through computer terminals at these institutions, as well as at Bad Arolsen. Currently, the digital archive contains approximately 190 million images of scanned documents.

1.2 The Central Name Index

Comprising almost 50 million cards and referencing an estimated 17.5 million individuals, the Central Name Index (CNI) serves as the primary finding aid for the ITS archive (Decker *et al.*, unpublished, p. 3). In the ITS digital archive, the scanned CNI cards serve as the primary finding aid for digital material. Because ITS was established as a tracing service, the majority of cards in the CNI reference individuals (Decker *et al.*, unpublished, p. 3). Accordingly, the CNI is indexed according to an alphabetic–phonetic system developed by ITS (Biedermann, 2007, p. 25). As described by the USHMM's CNI Card Guide, 'the CNI essentially operates like a physical library card catalog. Yet, whereas a card catalog references a book, only some CNI cards reference an original document' (Decker *et al.*, unpublished, p. 3). In this regard, the CNI is unlike a standard finding aid: some of the documents within the CNI are themselves original documents of historical interest.

The different types of CNI documents can be classified as follows, according to the USHMM's CNI Card Guide's taxonomy (Decker *et al.*, unpublished, p. 18):

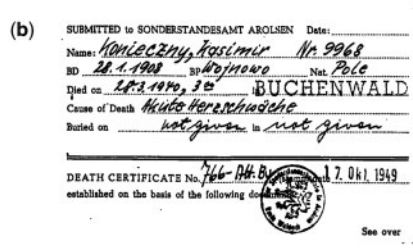
(1) Reference cards: linking the names of individuals to pertinent documents.
(2) Original cards, which are divided into three types:
   (a) Type 1: 'created by ITS or the International Refugee Organization (IRO) which do not reference another document' (Decker *et al.*, unpublished p. 57). Examples include:
      (i) Death certificate reference cards: documenting evidence provided by ITS for the issuance of death certificates by Sonderstandesamt Arolsen. These death certificates serve as official death records for a fraction of those who were confirmed to have perished within concentration camps (Christian Groh 2018, personal communication, 2 May).
      (ii) Inquiry cards: created 'in response to inquiry letters submitted by a third party' (Decker *et al.*, unpublished, p. 18).
      (iii) Internal hint cards: created to direct a researcher to other sources of information to consult (e.g. pointers to another name or card) (Decker *et al.*, unpublished, p. 50).
   (b) Type 2: 'Photocopies of original documents of which no other version of this document exists' (Decker *et al.*, unpublished, p. 57).
   (c) Type 3: 'Photocopies of original documents of which the original document does exist' (Decker *et al.*, unpublished, p. 57).

In Fig. 1, I present examples of thirteen different card types, classified according to this taxonomy; these thirteen examples represent only a small fraction of the diversity of cards within the CNI.

1.3 Sonderstandesamt Arolsen

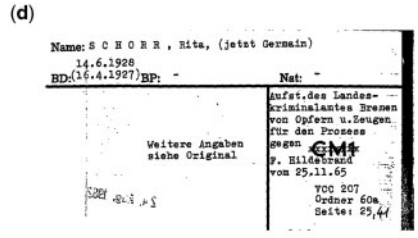
In the aftermath of the Holocaust, survivors and victims needed official documentation attesting to the deaths of family members for the purposes of legal claims such as restitution and compensation (Biedermann, 2007, p. 32). Sonderstandesamt Arolsen, the Special Registry Office in Bad Arolsen, Germany, was created on 1 September 1949, for the sole purpose of documenting the deaths of those who perished within concentration camps (Wittamer, 1950; Biedermann, 2007, p. 32). The official documentation produced by Sonderstandesamt Arolsen is a form of proof that is recognized for restitution and compensation purposes. The Sonderstandesamt is located near the physical ITS archive, which serves as its main source of information (HStAM Fonds 926 -

Sonderstandesamt Arolsen Death Certificate Reference Card, Variant 1 (Original Card Type 1) (CNI card of Zdenek Konecny, 0.1/28002393/ITS Digital Archive, USHMM)



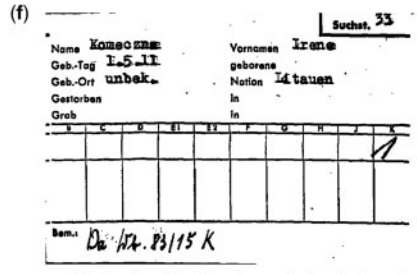
Sonderstandesamt Arolsen Death Certificate Reference Card, Variant 2 (Original Card Type 1) (CNI card of Kasimir Konieczny, 0.1/28004756/ITS Digital Archive, USHMM)

Sonderstandesamt Arolsen Death Certificate Reference Card, Variant 3 (Original Card Type 1) (CNI card of Margarete Konieczny, 0.1/28004353/ITS Digital Archive, USHMM)



T-Line Card (Reference Card) (CNI card of Rita Schorr, 0.1/38326699/ITS Digital Archive, USHMM)

Suchstelle (Tracing Bureau) Card, Hesse Variant (Reference Card) (CNI card of Jozef Konieczuy, 0.1/28005273/ITS Digital Archive, USHMM)



Suchstelle (Tracing Bureau) Card, Bavaria Variant (Reference Card) (CNI card of Irene Komeczna, 0.1/28004926/ITS Digital Archive, USHMM)

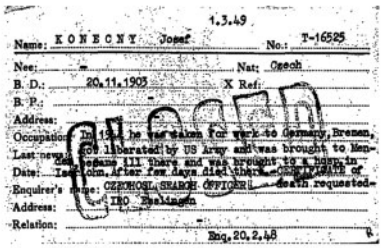
Fig. 1 Examples of thirteen different card types from the CNI, as classified according to the taxonomy of the USHMM's CNI Card Guide. Note the diversity of form structure and layout on the cards. Image (d) shows a CNI card for Rita Schorr, my grandmother, referencing material from another ITS subcollection concerning a trial in which she testified

Arcinsys Hessen, 2017). As of the Year 2000, the Sonderstandesamt had recorded more than 400,000 deaths, adding approximately 5,000–6,000 new entries every year (Maruhn, 2002, p. 233). In accordance with the German Civil Status Act of 19 February 2007, this certification process is exclusive to Sonderstandesamt Arolsen in Germany (Maruhn, 2002, p. 233). A translation of the pertinent section of the Civil Status Act is reproduced here (Personenstandsgesetz, 2007):

§38 Deaths in former concentration camps:

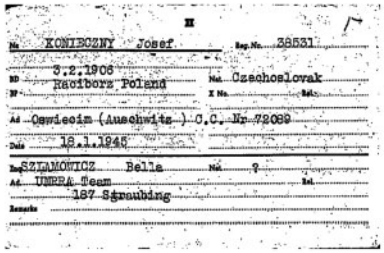
(1) The Special Registry Office in Bad Arolsen is exclusively responsible for the notarization of the deaths of prisoners of the former German concentration camps in Germany.

(g)



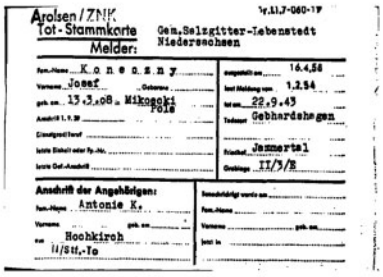
Inquiry Card, Variant 1 (CNI card of Josef Konecny, 0.1/28005270/ITS Digital Archive, USHMM)

(h)



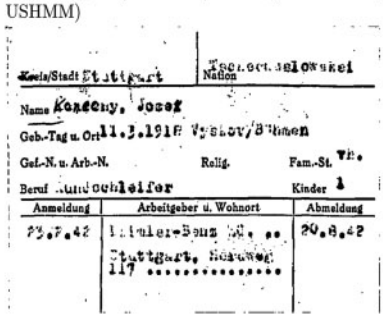
Inquiry Card, Variant 2 (CNI card of Josef Konieczny, 0.1/28005232/ITS Digital Archive, USHMM)

(i)



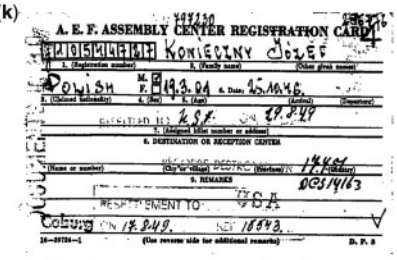
Zentralenamenskartei (ZNK) (Original Card Type 2) (CNI card of Josef Konecny, 0.1/28005215/ITS Digital Archive, USHMM)

(j)



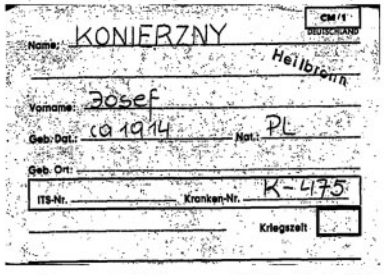
Württemberg-Baden Card (Original Card Type 2) (CNI card of Josef Konecny, 0.1/28005119/ITS Digital Archive, USHMM)

(k)



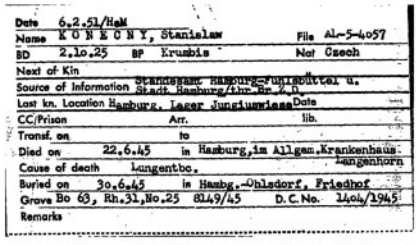
Allied Expeditionary Force (AEF) Registration Card (Original Card Type 2) (CNI card of Jozef Konieczny, 0.1/28005297/ITS Digital Archive, USHMM)

(l)



Care and Maintenance CM/1 Card (Reference Card) (CNI card of Josef Konierzny, 0.1/28005160/ITS Digital Archive, USHMM)

(m)



Multi-Line Card (Reference Card) (CNI card of Stanislaw Konecny, 0.1/28003281/ITS Digital Archive, USHMM)

Fig. 1 Continued

(2) The notarization of the deaths takes place on written notification of the certificate examination authority at the Special Registry Office in Bad Arolsen or the German office for the notification of the next of kin of the fallen of the former German Wehrmacht. The notification may also be filed by any person who was present at the time of death or informed of the death of his own knowledge. §3 (2) Sentences 1 and 4 and §4 (1) do not apply.[1]
(3) The notarization shall not take place if the death has already been recorded by another registry office. If documents cannot be obtained from this registry office, the death must be re-certified.

It must be stressed that the death certificates issued by Sonderstandesamt Arolsen do not document all prisoners who died within concentration camps; rather, they have been produced only for those who were confirmed to have died within the concentration camps according to surviving documentation. Because the primary surviving documentation used by Sonderstandesamt Arolsen was produced by the perpetrators, these death certificate reference cards are of historiographic interest to Holocaust scholars: by providing a demographic lens into prisoner deaths that were preferentially recorded by the perpetrators, these cards raise the important question of why these deaths and not others were recorded. Furthermore, understanding the systematic falsification of information on the cards, such as causes of death, reveals the intent of the perpetrators to mitigate and legitimize their genocidal actions.

1.4 Death certificate reference cards

Each death certificate reference card within the CNI is an ITS-produced card attesting to evidence provided by ITS and given to Sonderstandesamt Arolsen for the issuance of a death certificate (Christian Groh 2018, personal communication, 2 May). Because ITS is not the sole source of information for Sonderstandesamt Arolsen, not every death certificate issued by Sonderstandesamt Arolsen has a corresponding reference card within the CNI (Christian Groh 2018, personal communication, 2 May). However, because ITS is Sonderstandesamt Arolsen's primary source of information, these reference cards do exist for an appreciable fraction of death certificates.

I have identified four variants of death certificate reference cards within the digital CNI; examples of the fronts and backs of the four variants are depicted in Figs 2–5. The reference cards document demographic information, such as name, birth date, birthplace, nationality, death date, and death location. On the back of each card is a checklist noting the ITS documentation corroborating the information given to Sonderstandesamt Arolsen that appears on the front of the reference card.

In 2013, the Hessian State Archive Marburg received from Sonderstandesamt Arolsen a collection of 732 death registers, corresponding to death records on approximately 300,000 individuals; these death registers contain similar demographic information to the information found on the death certificates (HStAM Fonds 926 - Arcinsys Hessen, 2017). However, only 194 of these death registers have been digitized and made publicly accessible to date (LAGIS Hessen). Furthermore, there is currently no plan to digitize all of the death certificates (Christian Groh 2018, personal communication, 2 May). Consequently, the CNI contains the most comprehensive collection of digitized material pertaining to Sonderstandesamt Arolsen currently available to researchers. The question thus becomes whether these death certificate reference cards can be extracted from the digital CNI archive in an automated fashion.

1.5 Barriers to research

As described in (Lee, 2017), the digital CNI presents three primary challenges for the extraction of the death certificate reference cards (Lee, 2017, pp. 1–2):

(1) When the overwhelming majority of the CNI cards were scanned, the only consistent metadata recorded were the names of referenced individuals for indexing purposes. In other words, one can reliably search the digital CNI in a comprehensive fashion only by phonetic name and no other fields.
(2) Nearly all of the CNI cards were scanned as low-resolution binary TIFs.
(3) Handwritten information is prevalent on the overwhelming majority of the CNI cards.

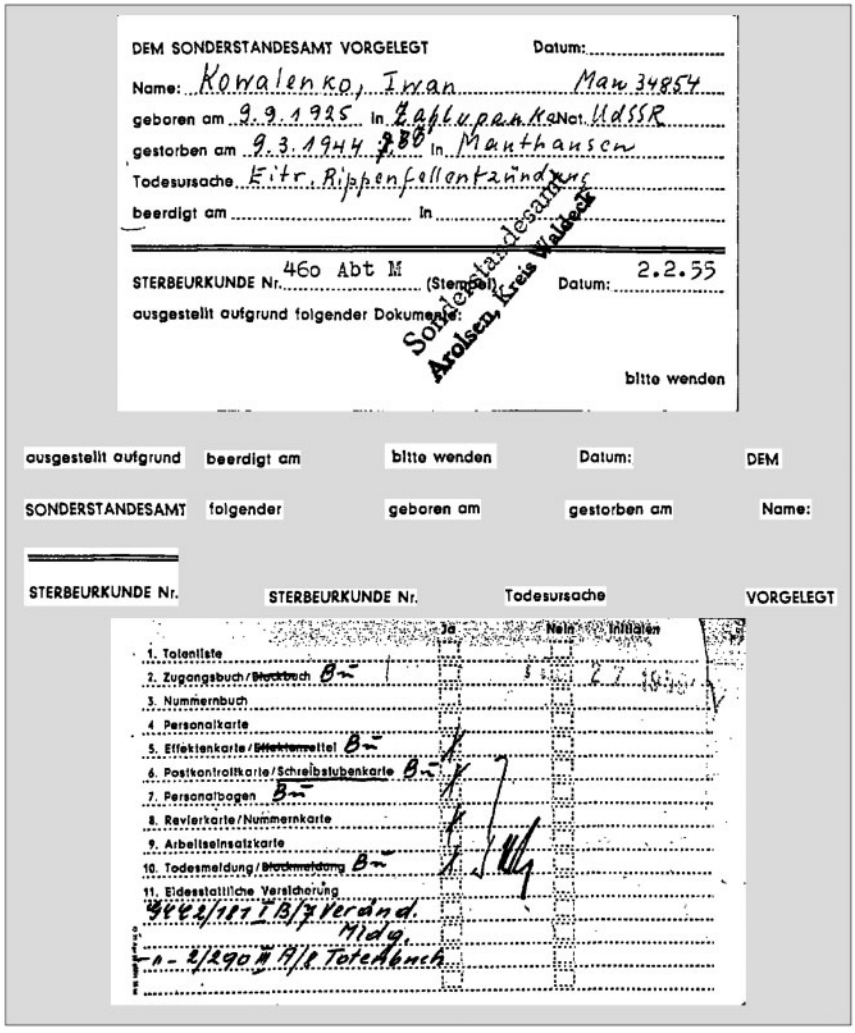


Fig. 2 The example image chosen to produce the templates for the two lines, German variant of death certificate reference card, as well as the fourteen templates extracted from the image used for template matching for this variant (CNI card of Iwan Kowalenko, 0.1/28494243/ITS Digital Arhive, USHMM). The back of the card is also shown

Ordinarily, the lack of metadata could be addressed by performing optical character recognition OCR to extract textual information amenable to searching. However, the second and third challenges preclude batch OCR from being effective. Consequently, alternative methods of extracting the death certificate reference cards must be considered. Because simple, automated checks, such as looking at image dimensions to isolate death certificates from other card types, do not perform well due to variances in image cropping, as well as distortions from the scanning process, more sophisticated methods must be utilized. In (Lee, 2017), I attempted to leverage the detection of line structure as a method of classification, which proved to be difficult due to the sensitivity of line detection to a number of different factors, including scan contrast and quality. This necessitated the search for alternative approaches.

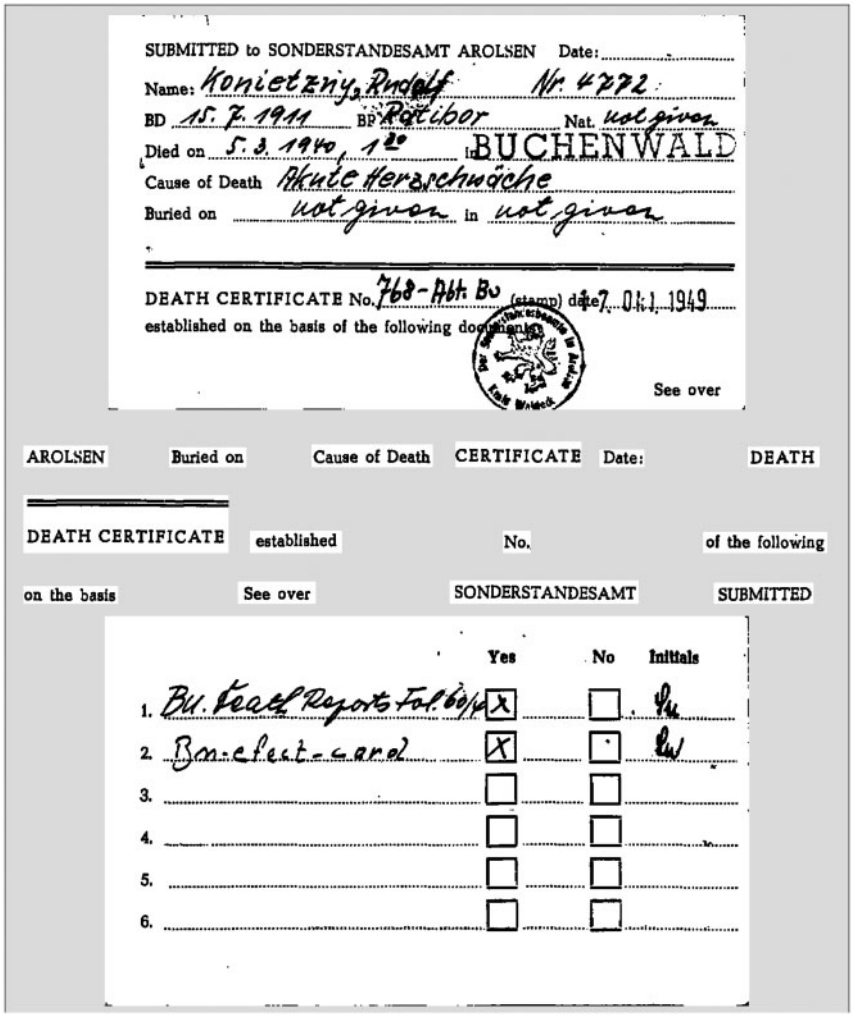


Fig. 3 The example image chosen to produce the templates for the two lines, English variant of death certificate reference card, as well as the fourteen templates extracted from the image used for template matching for this variant (CNI card of Rudolf Konietzny, 0.1/28003608/ITS Digital Archive, USHMM). The back of the card is also shown

1.6 Computer vision and machine learning

Recent advances in computer vision and machine learning, in conjunction with the exponential increase in computing power, have led to promising advances in training computers to perform image recognition and classification tasks. With the proliferation of software libraries such as `OpenCV` (Bradski, 2008), `scikit-learn` (Pedregosa *et al.*, 2011), and `TensorFlow` (Abadi et al., 2015), designed specifically for computationally efficient computer vision and machine learning routines, users are now able to train their own personal computers to perform many of these image classification tasks.

Indeed, the problem of extracting the death certificate reference cards from the CNI can be posed as an image classification problem: an image of a card can be classified either as any of the variants of a

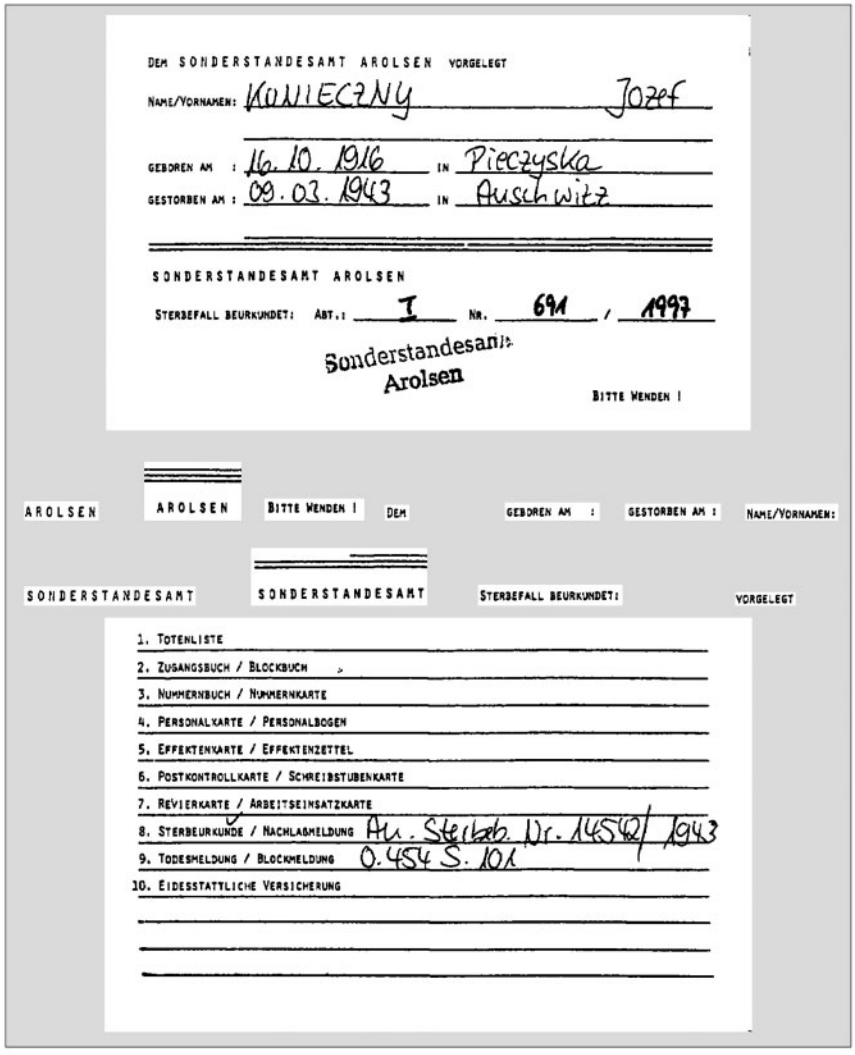


Fig. 4 The example image chosen to produce the templates for the three lines variant of death certificate reference card, as well as the eleven templates extracted from the image used for template matching for this variant (CNI card of Jozef Konieczny, 0.1/28005136/ITS Digital Archive, USHMM). The back of the card is also shown

death certificate reference card or as another card type (thus belonging to the negative class). Consequently, the question becomes whether it is possible to train a computer to automate this classification task. Fortunately, there are simple techniques from computer vision and machine learning that make this classification not only possible but computationally feasible on just a single personal computer.

1.6.1 *Template matching*

Template matching is a computer vision technique that entails comparing a pre-chosen template image with another, larger image to determine how much the template and regions of the image match. Template matching is performed by comparing the template to all regions of an image and evaluating the cross-correlation of the template and each corresponding image region, where cross-correlation is

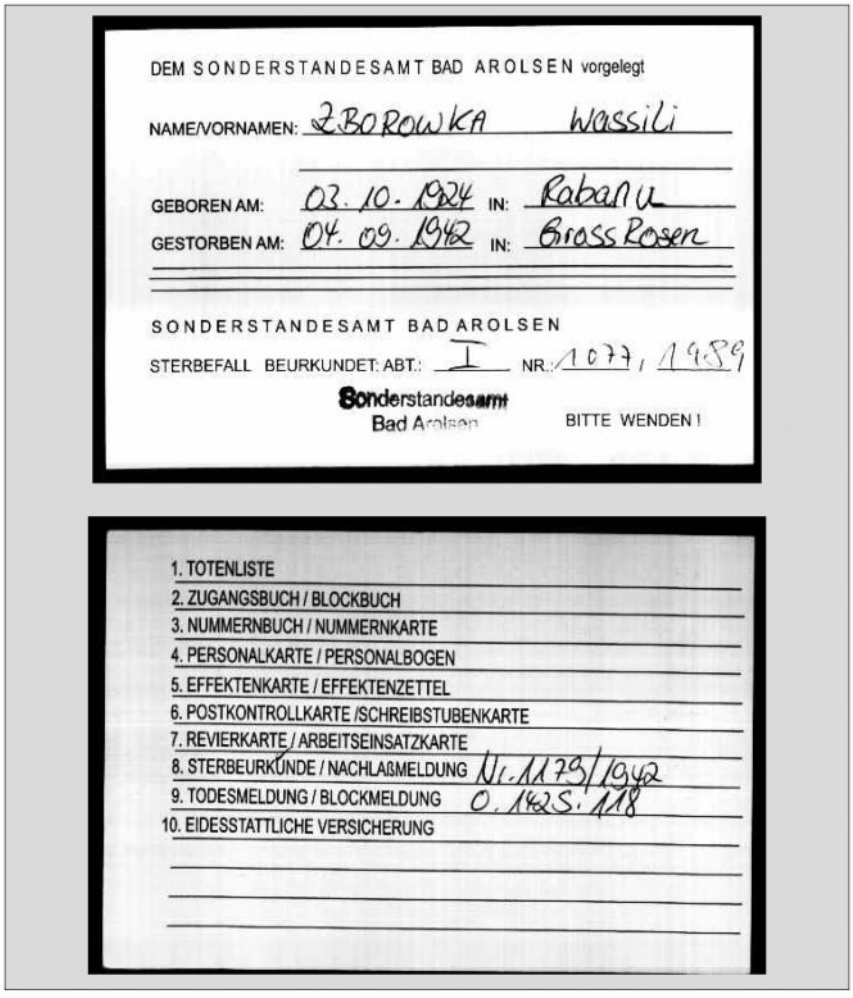


Fig. 5 An example of the front and back of the fourth variant of death certificate reference card (CNI card of Wassili Zborowka, 0.1/91360004/ITS Digital Archive, USHMM). To my knowledge, this variant exists only in JPG form

a metric for the pixel-wise similarity of two images. This yields a matrix of values corresponding to the cross-correlation of the template and each template-sized region on the image.[2]

Template matching is particularly well suited for the task of classifying death certificate reference cards within the CNI because this task requires identifying whether each card has a form structure consistent with any of the variants of the death certificate reference cards. In particular, one can select templates that capture the form structure of each variant of death certificate reference card and then perform template matching with each of these templates to assess the extent to which each template matches regions of each image.

Template matching provides a more robust approach for classification than the line detection method introduced in (Lee, 2017) because template matching directly leverages the form structure intrinsic to the cards.

1.6.2 *Supervised learning with machine learning classifiers*

The goal of classification with machine learning is to learn an algorithm that maps a set of input features to an output classification label according to a specified taxonomy. With supervised learning, a machine learning classifier is trained and evaluated using labeled data: examples with ground-truth classification labels that have been assigned by another means.

The canonical workflow with supervised learning is as follows. First, a set of labeled data is partitioned into a training set, validation set, and test set. A machine learning classifier uses the training set, including classification labels, to learn a classification function. The classifier is then fed the validation set without the corresponding ground-truth class labels and predicts the class labels according to its learned classification function. Based on the performance of the classifier, which is assessed by comparing the predicted labels to the ground-truth labels, the classifier's hyperparameters are tuned, and the classifier is subsequently re-trained with the training set, after which the performance on the validation set is re-assessed. Variants of this procedure include cross-validation; for example, k-fold cross-validation refers to a procedure in which the training/validation set is partitioned into k equal-sized folds, and classifier performance is assessed k times using one of the folds as a validation set and the remaining $k - 1$ folds as a training set. This process is repeated iteratively until the user has selected a final classifier with tuned hyperparameters. The test set is then utilized for the final evaluation of the classifier performance before using the classifier for the desired task of classifying unlabeled data. Unlike the validation set, the test set is withheld until after the classifier has been fully tuned to ensure that the performance metrics on the test set are not contaminated by subsequent user intervention via model tuning.

Two common metrics used to assess performance are precision and recall, which are adopted in this article as the chosen performance metrics. For binary classification, precision is defined as:

$$\text{precision} = \frac{TP}{TP + FP} \tag{1}$$

where TP is the number of true positives, and FP is the number of false positives. Similarly, recall is defined as:

$$\text{recall} = \frac{TP}{TP + FN} \tag{2}$$

where FN is the number of false negatives. Intuitively, precision is the fraction of data points assigned to a class that are true members of that class. Recall is the fraction of true members of a class that have been assigned to that class correctly.

Multi-class classification concerns the case in which there are more than two classes in the classification taxonomy. In this case, true positives, true negatives, false positives, and false negatives can be evaluated for each of the classes by reducing to a binary taxonomy for each class (e.g. for class i, does a data point fall into class i or not into class i?). Aggregate precision and recall values can then be calculated for n-class classification according to two different methods:

(1) Macro-averaging, i.e. calculating the precision and recall for each class and then averaging each metric across all classes:

$$\begin{aligned} \text{precision} &= \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FP_i} \\ \text{recall} &= \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FN_i} \end{aligned} \tag{3}$$

(2) Micro-averaging, i.e. calculating a global precision and a global recall:

$$\begin{aligned} \text{precision} &= \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i)} \\ \text{recall} &= \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FN_i)} \end{aligned} \tag{4}$$

In the death certificate reference card classification problem, the class labels correspond to the different

variants of the death certificate reference cards, as well as a negative class, and the input features for a given image can be taken to be salient values returned from template matching.

1.7 Ethical considerations

The application of machine learning and computer vision to Holocaust material raises a host of ethical questions. In the words of UCLA professor and digital humanities practitioner Todd Presner, 'Might . . . the realm of the "digital" and the "computational"—precisely, because it is, by definition, dependent on the algorithmic calculations, information processing, and discrete representations of data in digitized formats (such as numbers, letters, icons, and pixels)—present some kind of limit when it comes to responsible and ethical representations of the Holocaust?' (Presner, 2016, p. 179). Machine learning and computer vision are certainly not exempt from this line of questioning. Indeed, given the fraught debates that have surrounded the applications of these techniques to other domains, it is imperative that any application of these methods to Holocaust material is done ethically and with the utmost consideration of the victims.

I contend that the research presented in this article abides by these stipulations. First, distinction must be drawn between a document in an archive and the person described by the document: this research serves to analyze and classify CNI cards within ITS, not the human beings whom they describe. Second, because card type is an intrinsic property of a CNI card, the classification of CNI cards by type is a task with an *a priori* answer, unlike subjective tasks related to the document's content, such as sentiment analysis or topic modeling. Finally, though this article includes many figures and frequently quotes statistics describing the efficacy and computational efficiency of this method, the intent is neither to dehumanize nor to aestheticize the content of the archive; rather, these plots and statistics are necessary features of demonstrating the effectiveness of this method in improving access to the unique information and stories preserved within the archive.

2 Methodology

2.1 The digital copy of the CNI

I was given access to a file directory containing binary TIF images of the fronts and backs of 39,967,358 cards in the digital CNI. The scanned cards in this file directory represent 86% of the digital CNI, which presently comprises 46,423,197 cards; the additional 14% of scanned CNI cards not present in this file directory were not made available for this research but could in principle be extracted with database access. For this research, only the fronts of the CNI cards were processed, corresponding to 39,967,358 binary TIF images and 240 gigabytes of image data. It should be noted that this file directory also contained images of 96,629 cards (0.24%) in JPG form; however, the analysis in this article was restricted to TIFs, and these JPGs were omitted.

2.2 Computing resources and privacy considerations

Given the sensitive nature of the information preserved within the ITS digital archive, significant consideration was given as to where this file directory would be stored (McDonald, 2007, p. 1365). It was decided that these privacy considerations precluded using a third-party computing cluster to perform the classification and that the file directory would remain onsite on a machine at the USHMM. The onsite machine chosen for this research was a Velocity Micro Raptor^{TM}Signature Edition gaming PC with an 8-core Intel Core^{TM}i7-7820X X-series Processor and 32 GB of RAM. Though the use of only a single CPU placed significant constraints on the computational resources available for this research, it ultimately necessitated only slight modifications to the method, as described in Section 2.5.

I wrote all code in Python. I also utilized the computer vision library `OpenCV` (Bradski, 2008), the machine learning library `scikit-learn` (Pedregosa *et al.*, 2011), the scientific computing library `Numpy` (Oliphant, 2006), and the plotting library `Matplotlib` (Hunter, 2007).

2.3 Constructing the training, validation, and test sets: Taxonomy and beyond

To create labeled data, I classified a total of 32,865 CNI cards by hand. I adopted the following taxonomy for the classification:

(1) Death certificate reference card (two lines, German)
(2) Death certificate reference card (two lines, English)
(3) Death certificate reference card (three lines)
(4) Miscellaneous (the negative class)

It should be noted that, to my knowledge, the fourth variant of death certificate reference card, depicted in Fig. 5, exists only in JPG form. Because only binary TIFs were considered for classification in this article, this fourth variant was omitted from the classification taxonomy and is left for future classification with the remaining 96,629 cards in the file directory (0.24%) scanned as JPGs.

For the training/validation set,[3] I hand-classified 10,423 cards drawn randomly from this file directory; the breakdown of these initial, randomly drawn cards can be found in Table 1. Upon inspection of the table, it is immediately apparent that there is a strong class imbalance. To address the variance associated with having so few death certificate reference cards, I manually downloaded all TIFs of the death certificate reference cards that were returned with a keyword search of Sonderstandesamt in OuSArchiv, the user interface for accessing the ITS digital archive; this resulted in an additional 325 death certificate reference cards, as detailed in Table 1. These cards were added to the training/validation set, producing a training/validation set of 10,748 cards in total. For the test set, I hand-classified 22,117 cards, as detailed in Table 1. The cards in the test set were randomly drawn from the CNI and not injected with additional death certificate reference cards to retain the proportions found within the CNI as closely as possible.

2.4 Pre-processing uncropped images

Of the 39,967,358 images in the file directory, 1,739,494 images (4.4%) were not cropped during the scanning process; all such images are of the entire 1,376 pixel × 1,024 pixel scanning bed. These uncropped scans pose difficulties for template matching: in some instances, the cards are rotated relative to the scanning bed, and template matching is not robust enough to account for rotation.[4] Consequently, I introduced a simple automated cropping function to handle these scans. The pseudocode for this function is given in Algorithm 1.

Algorithm 1: Pseudocode for a function that attempts to crop an uncropped scan

```
Function crop_image (uncropped image)
  find image contours using OpenCV findContours();
  identify the largest contour on the image;
  find the minimum area rectangle containing
  this contour using OpenCV minAreaRect();
  crop and rotate image to horizontal according
  to this rectangle;
  if cropped image dimensions exceed (800, 600)
  then
      return cropped image
  else
      return uncropped image
  end
```

The choice of the cropping failure criterion in Algorithm 1 (a lower bound of 800 pixels × 600 pixels for a successful crop) is justified as follows. Histograms of the dimensions of the proposed crops for the 328 uncropped scans in the training/validation set, which represent 3.0% of the set, are depicted in Fig. 6. The dimensions of the failed crops are distinct from the dimensions of the successful crops, with thresholds of 800 pixels for the width of the crop and 600 pixels for the height of the crop, as shown in Fig. 6. These failed crops are the results of two phenomena:

(1) Poor scan quality: If the scan is saturated, OpenCV cannot differentiate the card from the background, and the proposed crop is a very small region of the image that is coupled to scan noise.
(2) Cropping error: The bounding rectangle identified by OpenCV represents a smaller rectangular region on the card (for example, one of the smaller rectangles in a T-line card partitioned by the T, as pictured in (d) in Fig. 1).

Table 1 A table displaying the classification breakdown of the training/validation set and test set according to the adopted taxonomy

Card type	Training/validation set			Test set
	Random	Added death certificate	Combined	Random
Death certificate (two lines, German)	54 (0.52%)	271	325 (3.0%)	144 (0.65%)
Death certificate (two lines, English)	5 (0.048%)	26	31 (0.29%)	19 (0.09%)
Death certificate (three lines)	10 (0.10%)	28	38 (0.35%)	24 (0.11%)
Misc.	10,354	N/A	10,354	21,930
Total	10,423	325	10,748	22,117

Note: The reported percentages in a given column reflect the percentages relative to the total number reported in the bottom row of the table.

For the training/validation set, 50 of the 326 uncropped scans failed this cropping procedure, corresponding to 15% of the uncropped scans and 0.47% of the total images in the training/validation set. Of these fifty scans, half corresponded to scans with significant portions that were entirely unintelligible due to scan quality, and the other half corresponded to cropping failures. These fifty scans were automatically separated out from the training/validation set during this pre-processing step.

Of the 1,739,494 uncropped images within the CNI, 1,538,327 images, or 88%, were cropped successfully with this cropping procedure. Relative to the total number of scans, only 0.50% failed the cropping procedure. These failures were flagged and separated out from the template matching procedure to be processed later. In total, 39,766,191 files (99.5%) were sent to template matching and classification.

2.5 Template matching: Construction and execution

I constructed the templates by first selecting a representative image of each death certificate type and then manually cropping templates according to features of the form structure that I determined to be intrinsic to the card type.[5] The templates selected for each of the three death certificate reference card types are depicted in Figs 2–4. In summary, fourteen templates were chosen for the two lines, German variant; fourteen templates were chosen for the two lines, English variant; and eleven templates were chosen for the three lines variant, amounting to a total of thirty-nine templates.

For each template, the template matching was performed on a region of each card equal to twice the extent of the template's dimensions.[6] Each region was centered to within 10% of the template's location relative to the image from which it was cropped; if this region extended beyond the extent of the image, the region was cropped accordingly. This restricted area template matching was chosen primarily for runtime considerations: as described in Section 2.2, the classification had to be performed on a machine with a single, 8 core Intel Core^{TM}i7-7820X X-series Processor. Under the assumption that the area of each CNI card is constant (a reasonable approximation), the computational complexity of naive template matching in this case scales as:

$$\text{runtime} \propto \left[\text{Area(card)} \times N_{cards} \times \sum_{\text{t} \in \text{templates}} \text{Area(t)} \right]. \quad (5)$$

Thus, the runtime can be decreased by reducing the area of each card used for template matching.[7]

For each template and each image, template matching was performed, and a three-dimensional vector (x, y, max) was recorded: x and y correspond to the normalized coordinates of the maximum cross-correlation value, and max is the maximum value itself. Thus, for each image, a 39×3 array was recorded for the thirty-nine templates, representing the features that formed the 117-dimensional input vector for each image.[8]

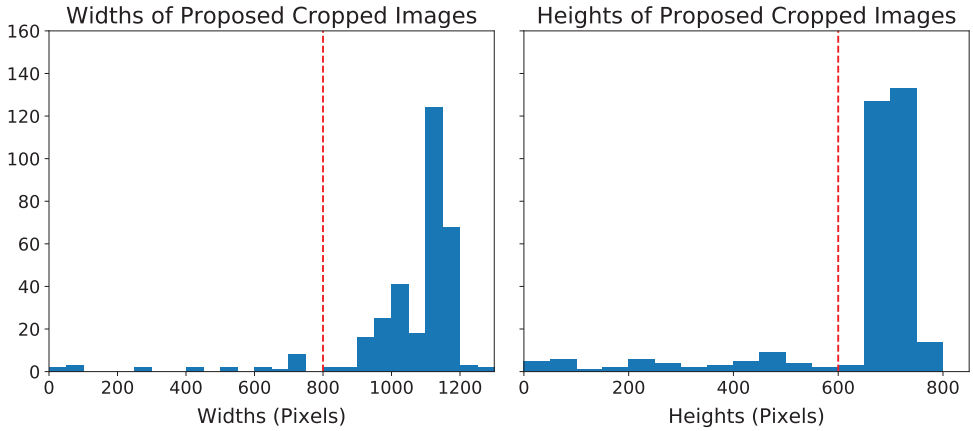


Fig. 6 Histograms of the widths and heights in pixels of the proposed crops for uncropped scans in the pre-processing step. The proposed threshold cuts of 800 pixels and 600 pixels for width and height, respectively, are shown as vertical dashed lines

2.6 Comparison of classifiers and feature selection

To compare classifier performance, I repeated five-fold cross-validation 100 times with random partitions of the training/validation set (five-fold cross-validation entails repeated partitioning into 80% training and 20% validation). As described in Section 2.4, 50 of the 10,748 images in the training/validation set failed the cropping procedure and thus were removed, leaving a training/validation set of 10,698 cards. All machine learning classification was performed using `scikit-learn` (Pedregosa *et al.*, 2011). I chose to compare the following classifiers:

(1) Naive Bayes
(2) Linear Support Vector Machine
(3) Random Forest (10 trees)
(4) Random Forest (100 trees)

Unless otherwise noted, the default hyperparameter values for the classifier as given by `scikit-learn` were used (Pedregosa *et al.*, 2011). In particular, I compared the four classifiers under the following classification conditions:

(1) One versus One classification compared to One versus Rest classification (two different methods for extending binary classifiers to the general multi-class case)
(2) 117 input features per image (corresponding to cross-correlation maxima of the cross-correlation as well as their x and y coordinates for the thirty-nine templates, as described in Section 2.5) versus thirty-nine input features (corresponding to just the cross-correlation maxima)

In Table 2, I present a comparison of the classifiers under the different classification conditions according to the metrics of precision and recall. Three phenomena are observed upon inspection of this table. First, One versus Rest classification consistently outperforms One versus One classification. Second, the Random Forest classifier with 100 trees consistently performs best among the different classifiers, according to both precision and recall. Third, the 39-feature input vector of just cross-correlation maxima consistently outperforms the 117-feature input vector. Indeed, when consulting the feature importances returned by the Random Forest classifier, the cross-correlation maxima were consistently ranked as more important than any of the features corresponding to x or y coordinates; by eliminating the features of lower importance, it is probable that overfitting was reduced.[9] As indicated in bold in the Table, the best performing classifier is the Random Forest with 100 trees, One versus Rest classification, and thirty-nine input features. It should be noted that the results in the Table have overlapping confidence intervals, but it is reasonable to believe that choosing the maximally performing

Table 2 A comparison of classifiers under different conditions, using the metrics of precision and recall

Classifier	One versus One		One versus Rest	
	Precision (SD)	**Recall (SD)**	**Precision (SD)**	**Recall (SD)**
117 input features (*x*, *y*, *max* for each template)				
Naive Bayes	99.95% (0.072%)	97.49% (2.2%)	99.94% (0.11%)	97.78% (2.2%)
Linear SVM	99.90% (0.17%)	98.49% (2.0%)	99.96% (0.089%)	99.30% (1.1%)
Random Forest (10 trees)	99.90% (0.27%)	98.91% (2.0%)	99.98% (0.056%)	99.38% (1.1%)
Random Forest (100 trees)	99.99% (0.046%)	99.58% (0.68%)	99.99% (0.031%)	99.64% (0.45%)
39 input features (only *max* for each template)				
Naive Bayes	99.09% (0.52%)	98.52% (1.6%)	97.91% (0.76%)	98.65% (1.5%)
Linear SVM	99.97% (0.062%)	99.22% (0.93%)	99.97% (0.062%)	98.45% (1.9%)
Random Forest (10 trees)	99.89% (0.25%)	99.06% (1.7%)	99.98% (0.076%)	99.52% (1.2%)
Random Forest (100 trees)	99.99% (0.042%)	99.52% (1.0%)	**99.99% (0.0094%)**	**99.74% (0.30%)**

Note: These results were produced by performing five-fold cross-validation 100 times. In particular, precision and recall have been computed by macro-averaging the metrics across the four classes due to the strong class imbalance, and then averaging over all five-folds and all 100 realizations. Each value reported in parentheses is the standard deviation of each metric across the five-folds, pooled over all 100 realizations. The values in bold refer to the maximal values of precision and recall.

classifier according to these results is acceptable. Because the classifier performance was so high, hyperparameter optimization was omitted.

In Fig. 7, I present the aggregated confusion matrix for all 100 realizations of five-fold cross-validation using a Random Forest with 100 trees, One versus Rest classification, and thirty-nine input features. Over the 100 combined realizations, all misclassifications manifested as assigning two lines, German death certificate reference cards as members of the negative class. In particular, these 332 misclassifications were repeated misclassifications of the same six cards, which are depicted in Fig. 8. The classifications reflect the expected failure modes of template matching, which is not robust to the rotation or stretching of a template.

2.7 Test set performance of chosen classifier

With the classifier chosen, the test set was then utilized as a final assessment of classifier performance. In total, 181 of the 1,914 uncropped cards failed the cropping procedure detailed in Section 2.4 and thus were removed from the test set, leaving a test set of 21,936 cards (99.2%). Because the test set was created by drawing cards randomly, whereas the training/validation set was artificially injected with additional death certificates (as described in Section 2.3), it is reasonable to expect that the test set better represents the underlying distribution of the CNI. Consequently, utilizing a test set serves also to confirm the robustness of the classifier to a distributional shift. Running the classifier on the test set, I find that the classifier performs with 100% precision and 100% recall for all four classes, allaying fears of a distributional shift and suggesting that this classifier is performing as desired.

3 Results

3.1 Classification of 40 million scanned cards from the CNI

As described in Section 2.4, 39,766,191 images (99.5%) were sent to template matching and classification in total. Before the classification step, the training/validation set was combined with the test set to produce a larger training set and thus maximize the amount of available labeled data. In total, the combined steps of the cropping procedure and template matching required 13.75 days of elapsed real time utilizing 6 of the 8 cores of the CPU. The classification step required 35 min of elapsed real time with a single core.

Of the 39,766,191 CNI cards that were classified, 312,183 cards (0.79%) were identified as

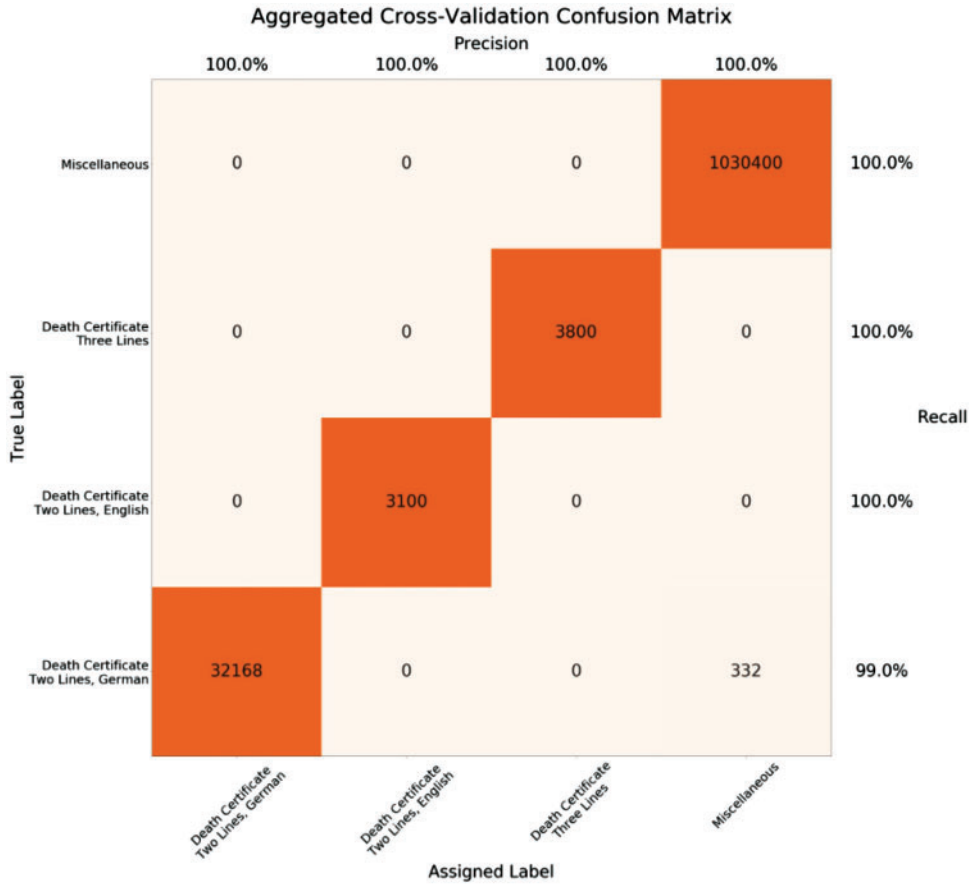


Fig. 7 An aggregated confusion matrix for 100 realizations of five-fold cross-validation. The precision and recall for each class are enumerated on the top and right of the confusion matrix, respectively

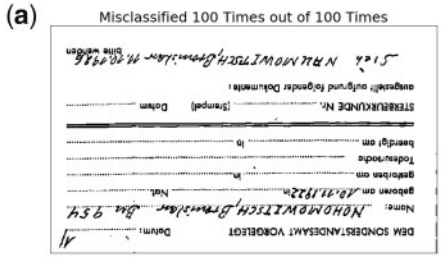
Sonderstandesamt Arolsen death certificate reference cards. The breakdown of these death certificate reference cards is as follows:

(1) Death certificate reference card (two lines, German): 255,567 cards (0.64%)
(2) Death certificate reference card (two lines, English): 32,962 cards (0.083%)
(3) Death certificate reference card (three lines): 23,654 cards (0.059%)

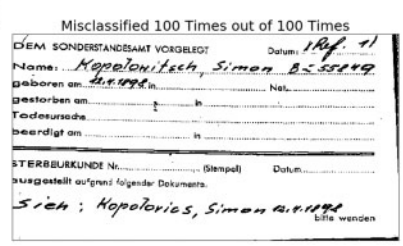
Given the nature of the cross-validation misclassifications presented in Section 2.3, it is reasonable to expect that virtually all of these 312,183 cards have been classified correctly, and death certificate reference cards that were misclassified as miscellaneous cards are symptomatic of the same failure modes depicted in Fig. 8.

Estimating that there are in total 480,000 Sonderstandesamt Arolsen death certificates according to the information presented in Section 1.3, the existence of 312,183 death certificate reference cards indicates that approximately two-thirds of all Sonderstandesamt Arolsen death certificates have been produced with ITS documentation.

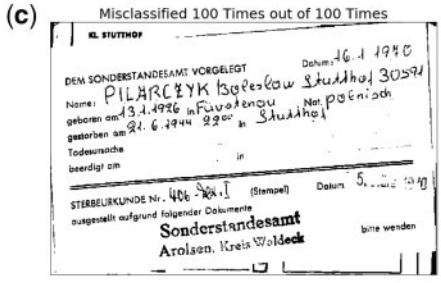
These death certificate reference cards have been copied to a new file directory and are now available to staff researchers at the USHMM, representing the first time that the cards have been aggregated together for research in digital form. In addition, they are currently being reincorporated into the digital



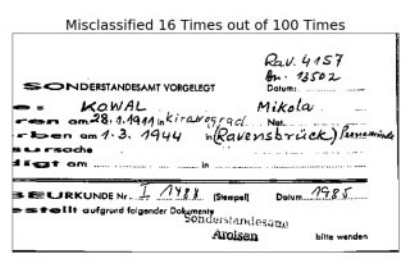
Misclassified 100% of the time due to its 180 degree rotation relative to the scanning bed (CNI card of Bronislaw Nohomowitsch, 0.1/56390786/ITS Digital Archive, USHMM)



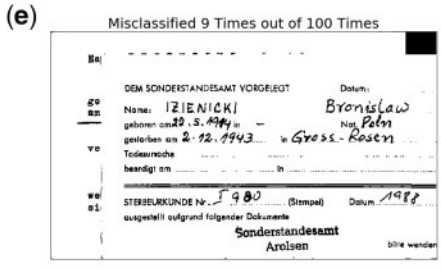
Misclassified 100% of the time due to horizontal stretching (CNI card of Simon Kopolowitsch, 0.1/62174890/ITS Digital Archive, USHMM)



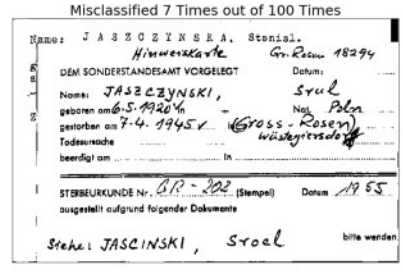
Misclassified 100% of the time due to the card's rotation relative to the scanning bed (CNI card of Boleslaw Pilarczyk, 0.1/62236476/ITS Digital Archive, USHMM)



Misclassified 16% of the time due to horizontal stretching. Only a small portion is stretched, leading to better performance than for 8b (CNI card of Mikola Kowal, 0.1/126185862/ITS Digital Arhive, USHMM)



Misclassified 9% of the time, possibly due to the image being miscropped, coupled with restricted area template matching (CNI card of Bronislaw Izienicki, 0.1/63631590/ITS Digital Archive, USHMM)



Misclassified 7% of the time, possibly due to the image being miscropped, coupled with restricted area template matching (CNI card of Srul Jaszczynski, 0.1/64063631/ITS Digital Archive, USHMM)

Fig. 8 Plots of the six misclassified cards over 100 realizations of five-fold cross-validation using a Random Forest classifier with 100 trees, One versus Rest classification, and thirty-nine input features. The number of times each card was misclassified over all 100 realizations is reported above each card. Furthermore, conjectured reasons for the misclassification are given in each caption

ITS archive as a new subcollection to be distributed to all ITS copyholders.

3.2 Toward a richer taxonomy

The taxonomy adopted throughout this article thus far consists of four classes: three classes of death certificate reference cards and a negative class. This taxonomy focusing on the death certificate reference cards was selected primarily because adding additional card types would require templates for each card type, which would in turn lead to an increase in runtime; this was avoided due to the constraint of being able to use only a single PC for this research, which would have rendered classification with a richer taxonomy intractable. However, it is nonetheless instructive to explore an expanded taxonomy according to the USHMM card guide, not only for testing the generalizability of the methodology but also for improving access to the CNI: as described in Section 1.2, there exists an abundance of other original documents in the CNI that do not exist anywhere else within ITS, including Siemens employee cards, displaced persons registration cards, and German Red Cross Zentralenamenskartei. Extracting these diverse documents would uncover a wealth of information for researchers.

In this section, I explore an expanded taxonomy of fourteen different classes. Representative examples of thirteen selected card types, classified according to document layout and form structure, are presented in Fig. 1 (the 14th class is the negative class). As a training/validation set, I use the same 10,698 card training/validation set used for classifier selection in Section 2.6 with updated labels for the fourteen classes. Because this expanded taxonomy was not intended to be applied to the full CNI (where runtime considerations are necessary), template matching was performed across the entirety of each card, in comparison to the restricted area template matching described in Section 2.5.

In Fig. 9, I report the aggregate confusion matrix for ten realizations of five-fold cross-validation. The macro-averaged precision is 99.6%, and the macro-averaged recall is 98.1%, indicating that this method generalizes well to a richer taxonomy for the CNI. It is worth noting that most misclassifications manifest as misclassifications with the negative class. The high performance of the classifier in the generalized taxonomy shows great promise for future work with ITS and other archives.

4 Future Work

There remains much to explore with the classification of the death certificate reference cards, as well as with all card types within the CNI. First, 14% of the digital CNI remains to be searched for death certificate reference cards. In addition, it would be beneficial to develop a more sophisticated cropping procedure for classifying the 201,167 scans (0.50%) that failed the cropping procedure; as described in Section 2.4, a fraction of these scans are unsalvageable due to scan quality, but the remaining scans are salvageable. Furthermore, it would be beneficial to extend the classification to JPG images because a small fraction (0.24%) of the 40 million scanned CNI cards available for this research were scanned as JPGs, as described in Section 2.1.

There are indeed other areas left to investigate with the classification of death certificate reference cards, in terms of both performance and runtime, the latter of which would be important if the taxonomy were to be expanded in accordance with Section 3.2 and applied to the entire CNI. In regard to performance, the cross-validation misclassifications presented in Fig. 8 were all due to intrinsic flaws of basic template matching: sensitivity to image rotation and scaling. Other methods, such as feature matching with SIFT, are both rotation and scale invariant, making them attractive options for overcoming these limitations of template matching (Lowe, 2004).

In principle, one could explore other more sophisticated methods such as neural networks; however, given the computational efficiency of my method and its near-perfect performance, these alternative approaches are left for future work. In addition, with the current implementation of template matching, it may be of use to experiment with smoothed templates or averaged templates, produced by averaging the same region of multiple cards of the same card type, to eliminate aberrancies in the templates,

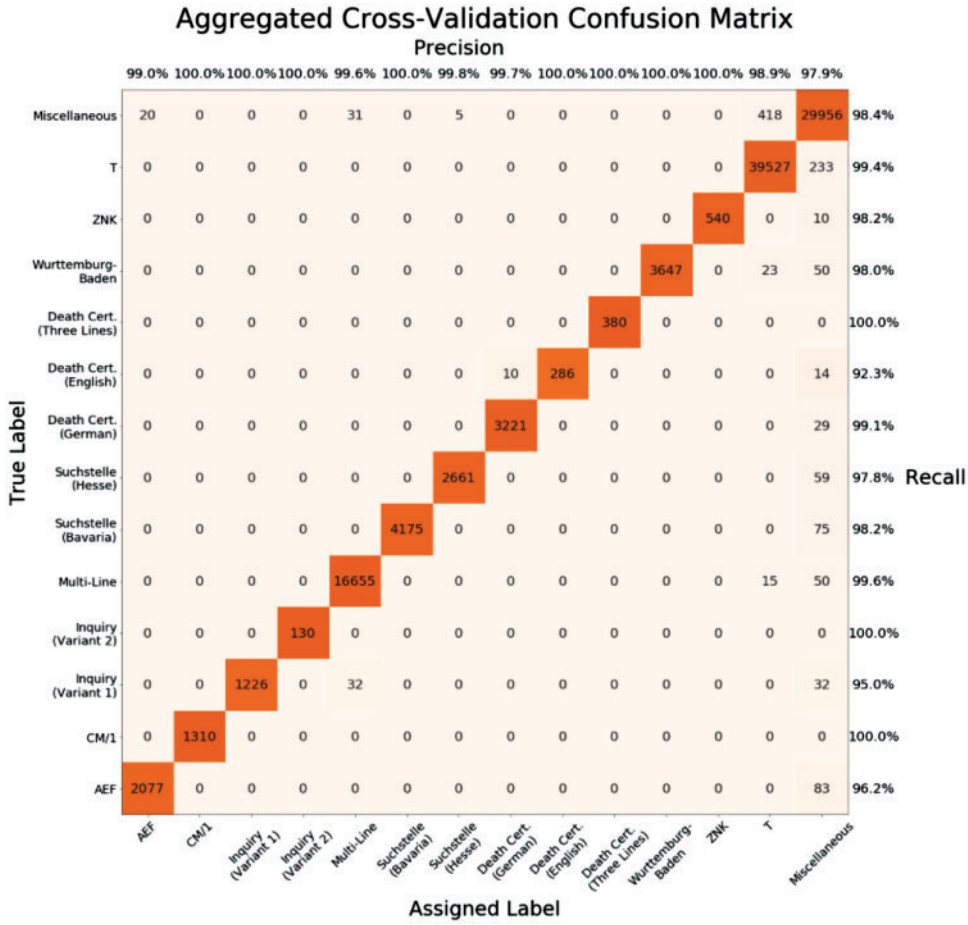


Fig. 9 An aggregate confusion matrix for ten realizations of five-fold cross-validation with the fourteen-class taxonomy. The precisions and recalls for each class are enumerated on the top and right of the confusion matrix, respectively

such as scan noise. In regard to runtime, improvements could be made by downsampling the images, optimizing the code, and exploring alternative methods such as feature matching.

Finally, given the sheer number of death certificates, it would be beneficial to explore the use of handwriting recognition to create a database of demographic information from the death certificate reference cards for historical research. Indeed, much progress has been made with handwriting recognition over the past few years utilizing long short-term memory and recurrent neural networks (Graves & Schmidhuber, 2009; Doetsch *et al.*, 2014; Pham *et al.*, 2014).

5 Conclusion

I have introduced an automated method that utilizes template matching and machine learning to extract reference cards to Sonderstandesamt Arolsen death certificates from 40 million images in the ITS digital archive. I have demonstrated the efficacy of this method by confirming near-perfect precision and recall on validation and test sets. Applying this method to 40 million scanned cards in the CNI, I successfully extracted 312,183 death certificate reference cards in 13.75 days of wall-clock runtime using a single, $600 processor. Finally, I showed that this method successfully generalizes to a richer,

fourteen class taxonomy within the CNI. The generalizability of this method, in conjunction with its computational efficiency, indicates that this algorithm could be applied to a wide diversity of archives without the need for expensive computational resources.

Acknowledgements

The author would like to thank the United States Holocaust Memorial Museum for supporting the author's fellowship and research. In addition, Harvard University's History Department and the International Tracing Service have been integral to this research. The author would also like to thank the following individuals: Michael Haley Goldman, Robert Ehrenreich, Gabriel Pizzorno, Michael Levy, Elizabeth Anthony, Diane Afoumado, Sara-Joelle Clark, Jo-Ellyn Decker, Laura Ivanov, Jude Richter, Paul Shapiro, Wolfgang Schneider, Christian Groh, Stephen Portillo, Saahil Mehta, and Mengting Zhang for their generous help throughout this research.

Funding

This work was supported by the United States Holocaust Memorial Museum.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale Machine Learning on Heterogeneous Systems. Software available from: tensorflow.org.

Agreement on the International Tracing Service. Agreement on the International Tracing Service: Partnership agreement on relations between the federal archives of the federal republic of Germany and the international tracing service, 2013. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/190234/Misc.1.2013.AgreementTracingService.pdf (accessed 1 February 2017).

Belkin, P. (2007). Opening of the international tracing service's holocaust-era archives in Bad Arolsen, Germany. *CRS Report for Congress*. https://fas.org/sgp/crs/misc/RS22638.pdf (accessed 21 February 2017).

Biedermann, C. (2007). *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*. Bad Arolsen: Wildner-Druck.

Bradski, G. (2008). The OpenCV Library. Dr. Dobb's Journal of Software Tools.

Brown-Fleming, S. (2016). *Nazi Persecutions and Postwar Repercussions: The International Tracing Service Archive and Holocaust Research*. Lanham: Rowan & Littlefield.

CNI card of Bronislaw Izienicki, 0.1/63631590/ITS Digital Archive, USHMM.

CNI card of Srul Jaszczynski, 0.1/63631590/ITS Digital Archive, USHMM.

CNI card of Irene Komeczna, 0.1/28004926/ITS Digital Archive, USHMM.

CNI card of Josef Konecny, 0.1/28005119/ITS Digital Archive, USHMM.

CNI card of Josef Konecny, 0.1/28005270/ITS Digital Archive, USHMM.

CNI card of Stanislaw Konecny, 0.1/28003281/ITS Digital Archive, USHMM.

CNI card of Zdenek Konecny, 0.1/28002393/ITS Digital Archive, USHMM.

CNI card of Josef Koneczny, 0.1/28005215/ITS Digital Archive, USHMM.

CNI card of Josef Konieczny, 0.1/28005297/ITS Digital Archive, USHMM.

CNI card of Josef Konieczny, 0.1/28005232/ITS Digital Archive, USHMM.

CNI card of Jozef Konieczny, 0.1/28005136/ITS Digital Archive, USHMM.

CNI card of Kasimir Konieczny, 0.1/28004756/ITS Digital Archive, USHMM.

CNI card of Margarete Konieczny, 0.1/28004353/ITS Digital Archive, USHMM.

CNI card of Jozef Konieczuy, 0.1/28005273/ITS Digital Archive, USHMM.

CNI card of Josef Konierzny, 0.1/28005160/ITS Digital Archive, USHMM.

CNI card of Rudolf Konietzny, 0.1/28003608/ITS Digital Archive, USHMM.

CNI card of Simon Kopolowitsch, 0.1/62174890/ITS Digital Archive, USHMM.

CNI card of Mikola Kowal, 0.1/126185862/ITS Digital Archive, USHMM.

CNI card of Iwan Kowalenko, 0.1/28494243/ITS Digital Archive, USHMM.

CNI card of Bronislaw Nohomowitsch, 0.1/56390786/ITS Digital Archive, USHMM.

CNI card of Boleslaw Pilarczyk, 0.1/62236476/ITS Digital Archive, USHMM.

CNI card of Rita Schorr, 0.1/38326699/ITS Digital Archive, USHMM.

CNI card of Wassili Zborowka, 0.1/91360004/ITS Digital Archive, USHMM.

Decker, J., Clark, S., and Ivanov, L. (Unpublished). *CNI Cards: The Holocaust Survivors and Victims Resource Center, the United States Holocaust Memorial Museum*.

Doetsch, P., Kozielski, M., and Ney, H. (2014). Fast and robust training of recurrent neural networks for offline handwriting recognition. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, Hersonissos, September 2014, pp. 279–84. https://ieeexplore.ieee.org/abstract/document/6981033/ (accessed 1 April 2018).

Graves, A. and Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in Neural Information Processing Systems 21*, Vancouver, December 2008. https://dl.acm.org/citation.cfm?id=2981848 (accessed 1 April 2018).

Hessische Geburten-, Ehe-, Sterberegister: Landesgeschichtliches Informationssystem Hessen. Hessischen Landesamts für geschichtliche Landes kunde (Marburg). https://www.lagis-hessen.de/de/subjects/gsearch/page/10/sn/pstr?q=bestand:926 (accessed 10 April 2018).

The Holocaust Survivors and Victims Resource Center at the United States Holocaust Memorial Museum (2007). *ITS Frequently Asked Questions*. https://www.ushmm.org/remember/the-holocaust-survivors-and-victims-resource-center/international-tracing-service/about-the-international-tracing-service/its-frequently-asked-questions (accessed 5 October 2017).

HStAM Fonds 926 - Arcinsys Hessen. (2017). Hessisches Staatsarchiv Marburg. https://arcinsys.hessen.de/arcinsys/detailAction.action?detailid=b6744 (accessed 10 April 2018).

Hunter, J. (2007). A 2D graphics environment. *Computing in Science and Engineering*, **9**: 90–5. (accessed 1 April 2018).

Lee, B. (2017). Line detection in binary document scans: a case study with the International Tracing Service archives. In *2017 IEEE International Conference on Big Data*, Boston, December 2017, pp. 2256–61. https://ieeexplore.ieee.org/document/8258178/ (accessed 10 February 2018).

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, **60**(2): 91–110.

Maruhn, S. (2002). *Staatsdiener im Unrechtsstaat: Die deutschen Standesbeamten und ihr Verband unter dem Nationalsozialismus*. Verlag für Standesamtswesen, Frankfurt am Main.

McDonald, C. (2007). Reconciling holocaust scholarship and personal data protection: facilitating access to the international tracing service archive. *Fordham International Law Journal*, **30**(4): 1360–91.

Oliphant, T. (2006). *A Guide to NumPy*. Trellgol Publishing.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, **12**: 2825–30.

Personenstandsgesetz vom 19. Februar (2007) (BGBl. I S. 122), das zuletzt durch Artikel 2 Absatz 2 des Gesetzes vom 20. Juli 2017 (BGBl. I S. 2787) geändert worden ist. https://www.gesetze-im-internet.de/pstg/BJNR012210007.html (accessed 12 April 2018).

Pham, V., Bluche, H., Kermorvant, C., and Louradour, J. (2014). Dropout improves recurrent neural networks for handwriting recognition. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, Hersonissos, September 2014, pp. 285–90. https://ieeexplore.ieee.org/abstract/document/6981034/ (accessed 1 April 2018).

Presner T. (2016). The ethics of the algorithm: close and distant listening to the shoah foundation visual history archive. In Fogu, C., Kansteiner, W. and Presner, P. (eds), *Probing the Ethics of Holocaust*

Culture. Cambridge: Harvard University Press, pp. 175–202.

Senate Resolution 142 – 110th Congress. (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. https://www.govtrack.us/congress/bills/110/sres142 (accessed 21 February 2017).

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**: 2579–605.

Wittamer, A. J. (1950). Early letter regarding Sonderstandesamt Arolsen, 6.1.1/82508787/ITS Digital Archive, USHMM.

Notes

1 According to §38 (2), the following sentences do not apply: §3 (2) Sentences 1 and 4: 'The civil status registers are kept electronically.... The program must allow an automated search on the basis of the information to be included in the civil status registers; the registers must be able to be evaluated at any time according to annual entries' (Personenstandsgesetz, 2007). §4 (1): 'The notarizations in a civil status register are to be stored after their conclusion (§3 Sentence 2) in another electronic register (security register)' (Personenstandsgesetz, 2007).

2 To gain some intuition, 'Where's Waldo' is an example of a well-suited problem for template matching: taking the template to be an image of Waldo and the larger image to be the full illustration of the crowd, the maximum value of the resulting cross-correlation matrix corresponds to the location of Waldo in the illustration.

3 As will be described in Section 2.6, I use cross-validation during the classifier selection, meaning that there is no single partition between the training and validation sets; I thus adopt the convention of calling the collective set the training/validation set.

4 If the template is sufficiently rotated relative to the image orientation, a region that would ordinarily yield a high cross-correlation value will yield a low cross-correlation value.

5 Because all images had the same scan resolution, more complicated versions of template matching, such as multi-scale template matching, did not have to be considered.

6 To perform the template matching, I used the `OpenCV` function `matchTemplate()` with the parameter `method=CV_TM_CCOEFF_NORMED` (Bradski, 2008).

7 This provides even more motivation for pre-processing uncropped scans, as described in Section 2.4: performing restricted area template matching requires having consistent normalized coordinates for the cards in the images.

8 This 117-dimensional vector constrains the card type so well that unsupervised learning techniques could be applied here: a heuristic analysis of clustering in this 117-dimensional space using `t-SNE` reveals a strong clustering decomposition according to card type (van der Maaten and Hinton, 2008).

9 It should be noted that the further removal of features from the thirty-nine-feature input vector resulted in noticeable performance loss as assessed with cross-validation, thus discouraging the removal of templates to improve runtime.