Extracting Meaningful Data from *Decomposing Bodies*

Alison Langmead
Department of History of Art and Architecture
University of Pittsburgh
PA 15260, USA
adlangmead@pitt.edu

Paul Rodriguez
San Diego Supercomputing Center
University of California San Diego
CA 92093, USA
p4rodriguez@ucsd.edu

Sandeep Puthanveetil Satheesan
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
IL 61801, USA
sandeeps@illinois.edu

Alan Craig
XSEDE
P.O. Box 5020
IL 61801, USA
a-craig@illinois.edu

ABSTRACT

We present *Decomposing Bodies*, a digital humanities project that examines the late-19th-century system of anthropometrical measurement introduced in France by Alphonse Bertillon. "Bertillonnage," as this system is commonly known, was the first measurement-based, state-controlled system used for criminal identification. Currently, researchers resort to the tedious manual transcription in order to study the data on these cards in bulk. Here, we propose an end-to-end system for extracting handwritten text and numbers from scanned Bertillon cards in a semi-automated fashion and also the ability to browse through the original data and generated metadata using a web interface. The proposed system will enable historians and humanities researchers to study the data produced by the Bertillon system with much more ease than ever before. To the best of our knowledge, this is the first system that has tried to automate Bertillon card analysis through the application of existing handwritten digit and word recognition methods. We present our current results on performing document analysis on a selected set of scanned Bertillon cards from the Ohio State Reformatory and Ohio Penitentiary. We conclude with a few recommendations for increasing the likelihood of success for collaborations between Computer Science and Digital Humanities researchers.

CCS CONCEPTS

• **Applied computing** → **Document analysis**; *Media arts*; Document scanning; *Optical character recognition*; • **Social and professional topics** → History of computing;

KEYWORDS

Bertillon cards, Bertillonnage, handwritten text recognition, handwritten number recognition, OCR, document analysis, digital humanities, information management history, criminal identification history, visual and material culture studies

ACM Reference format:
Alison Langmead, Paul Rodriguez, Sandeep Puthanveetil Satheesan, and Alan Craig. 2017. Extracting Meaningful Data from *Decomposing Bodies*. In *Proceedings of Practice and Experience in Advanced Research Computing 2017, New Orleans, Louisiana, USA, July 09-13, 2017 (PEARC17),* 8 pages.
https://doi.org/10.1145/3093338.3093368

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
PEARC17, July 09-13, 2017, New Orleans, Louisiana, USA
© 2017 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.
ACM ISBN 978-1-4503-5272-7/17/07…$15.00
https://doi.org/10.1145/3093338.3093368

1 INTRODUCTION

Decomposing Bodies is a digital humanities project housed in the Visual Media Workshop in the Department of History of Art and Architecture at the University of Pittsburgh that is focused on the study of a set of historical prison records currently held by the Ohio History Connection in Columbus, Ohio. These records took part in the Bertillon system of anthropometric identification as it was implemented in the United States of America [3]. Many of the research goals of this project would clearly benefit from a complete transcription of these handwritten records, a task which is not only highly-computationally expensive, but one that involves much high-level system training and machine learning skills. It was for this reason that a collaboration between the Pittsburgh *Decomposing Bodies* team and the XSEDE/ECSS team was formed. In this paper, we first present some background material on Bertillonnage and the goals for the overarching *Decomposing Bodies* project, followed by a discussion of the specific collaboration created between the team from the University of Pittsburgh and the XSEDE/ECSS team[1]. We then conclude with not only the technical results of the collaboration, but with some discussion of, and reflection on, the workings of the collaboration both from the technologist's and the digital humanist's points-of-view.

[1]The background material on Bertillonnage presented in this paper is part of a larger collaborative research initiative taking place between the first author and Dr. Josh Ellenbogen, Department of History of Art and Architecture, University of Pittsburgh.

Figure 1: Bertillon Card (recto and verso) documenting Roy Foley, Ohio State Reformatory Prisoner No. 2490, October 11, 1905. Ohio History Connection, State Archives Series 1416AV.

2 THE *DECOMPOSING BODIES* PROJECT

2.1 Alphonse Bertillon's Measure of Man

Late in the nineteenth century, Alphonse Bertillon, the French criminologist, anthropologist, statistician, and inventor developed a system of criminal identification that sought to classify human beings on individual standardized cards[2]. Each card used a pre-established set of eleven anthropometrical measurements (such as height, length of left foot, and width of the skull) as an index for other identifying information about each criminal (such as the crime committed, their nationality, and two photographs) (see Fig. 1). This decomposition of the body into numerical measurements, Bertillon argued, would allow the French police to fight back against one of their most pressing issues: capturing repeat offenders. New laws had recently been put into place which levied increased penalties for recidivism, and with these new rules had come greater efforts on the part of the criminal population to evade arrest by lying about their identities and their past behavior. In the absence of a reliable individuation technology, however, the state was powerless to contradict their self-reported stories. Bertillon's system of anthropometric identification, now more commonly known as "Bertillonnage," allowed the French state to become capable, in the first systematic and self-consciously scientific way, of harnessing individual citizens to fixed identities, which could then be attached to permanent records of personhood and activity.

Prior to Bertillon's tenure at the Préfecture of Police, the French state had made haphazard efforts at creating identity records, including the creation of thousands of individual photographs of arrested criminals all filed by surname. This measure, however, proved ineffective as the criminals needed only to lie about their surname to subvert the effectiveness of the index. Bertillon's system did not rely on any such suspect-made statement. Instead, since he believed that the variation of the eleven individual bony traits he placed on the cards followed the binomial distribution (something newly

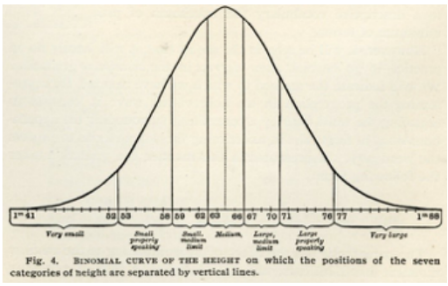

Fig. 4. BINOMIAL CURVE OF THE HEIGHT on which the positions of the seven categories of height are separated by vertical lines.

Figure 2: Binomial curve of distribution broken down into subgroupings. From Bertillon, Signaletic Instructions (Chicago, 1896), 38.

argued by fellow anthropologists such as Sir Francis Galton)[3], one could use these measurements to consistently break each measure down into sub-groupings of size that would encompass approximately equal numbers of cases (see Fig. 2). One could then organize the identity archive around the measurements, first dividing all the records into seven groups of head length, then subdividing those records by head width, then by left middle finger length, and so on until one ended up with small groupings that contained approximately a dozen cards. Storage worked like retrieval: take the prescribed eleven measurements from the body in question, then go to the corresponding section of the archive to find, or file, the appropriate card. These numbers thus served as an index not only for the person's name and photograph, but also for a large amount of socio-cultural data also stored on the card such as information about "Descent," "Crime," "Occupation," and "Color."

The success of this information retrieval system relied utterly on the ability to create, and closely repeat, the series of measurements needed to create the indices (keys) for the cards (values). In fact, producing a given measure in the Bertillon system required the bodies of both the police functionaries and the detained criminals to follow a choreography of motions that encompassed up to thirty specific steps. Bertillon himself argued that the measurers were supposed to be able to generate measurements that, for some values, would come within a half millimeter of anyone else who would ever apply the same system to the same person[4].

Although initially implemented at the Préfecture of Police in Paris, Bertillonnage spread internationally, coming to the United States in the late 1880s. At that time, R.W. McClaughry, the then Warden of the Illinois State Penitentiary at Joliet, introduced Bertillon's *Signaletic Instructions: Including the Theory and Practice of Anthropometrical Identification* to American prison administrators. The system then spread to several prisons in the U.S., including, significantly, Leavenworth, the first Federal Penitentiary.

[2]This system was most fully put forward in Alphonse Bertillon, *Identification anthropométrique; instructions signalétiques* (Melun: Imprimerie administrative, 1893). It was then translated into English under the title, *Signaletic Instructions: Including the Theory and Practice of Anthropometrical Identification*, edited by R. W. McClaughry (Chicago: The Werner Company, 1896).

[3]For the work of Francis Galton on the normal distribution in the context of natural selection, see his work *Natural Inheritance* (London: MacMillan, 1889). For further information on Galton and his relationship to Bertillon and other late-nineteenth century scientists, please see Josh Ellenbogen, *Reasoned and Unreasoned Images: The Photography of Bertillon, Galton, and Marey* (University Park, Pennsylvania: Pennsylvania State University Press, 2012).

[4]For the table of allowable errors in measurement, please see *Identification anthropométrique*, xxvi.

2.2 The Data and Research Agenda of *Decomposing Bodies*

The *Decomposing Bodies* project, housed in the Visual Media Workshop in the Department of History of Art and Architecture at the University of Pittsburgh, examines the Bertillon system in relation to a number of different areas of humanistic and social scientific inquiry. Before the age of digital machines, before the rampant quantification and standardization of the physical world were taken in stride, this practice of dissolving the body into numbers, images, and letters was novel, unknown. *Decomposing Bodies* seeks to defamiliarize this process of breaking down and defining what we see into quantified digests, by collecting, analyzing, digitizing, and re-presenting the data created by the process of Bertillonnage, specifically as practiced in the United States. It is for this reason that the collaboration was begun with XSEDE [12].

The Bertillon cards used in the *Decomposing Bodies* project are from two penal institutions in Ohio: the Ohio Penitentiary, founded in 1834 and located in Columbus, Ohio, and the Ohio State Reformatory, founded in 1886 and located in Mansfield, Ohio. The Reformatory, the newer of the two institutions, was created during a time of prison reform after the Civil War, and was an institution that placed greater emphasis on the rehabilitation and education of its inmates, most of whom tended to be younger men convicted of lesser crimes. The Penitentiary, on the other hand, was the location for the incarceration of the more hardened, older criminals. During this period, it had a long-standing reputation for being a harsh institution supported by the profits made by means of prison labor.

A team of researchers from the University of Pittsburgh's Visual Media Workshop has captured the images of over 12,500 cards from these penal institutions, all of which are now held by the Ohio History Connection in Columbus, Ohio. In order to achieve many of our research goals, we need to transcribe the data on these cards into a machine-readable format. Using these transcriptions, we envisage a number of research outcomes, some of which have already come to pass. For example, we have already produced an exhibition using themes from this research in a show entitled *Data (after)Lives*, held at the University Art Gallery at the University of Pittsburgh in the fall of 2016 [1]. In this exhibition, we created a digital media project using some of the images created by the XSEDE team, and also a series of objects made from the extracted images of the faces from the cards, created using the OpenFace facial recognition system [4].

For the near future, we also hope to use the transcribed data from these cards to produce a digital project that will focus on the relationships between identity and bodily measurement at the turn of both the twentieth century, using Bertillonnage as a focus, and the turn of the twenty-first century, using computer vision algorithms for facial and gait recognition as a focus. We plan to produce an interactive resource that investigates—using both text and images—the ways in which these two methods of rendering the human body into code participate in a complex dance of identity, control, and technology. Ultimately, by looking at what unites and divides these two technologies, we hope to illuminate some of the different ways in which individuals have interacted with state power in modernity, and how their experiences of it have altered through time.

Finally, we hope to produce research products that will bring out the socio-historical and race-based implications of Bertillonnage, especially as practiced here in the United States. Since the records at the Ohio History Connection span several decades, we will be able to investigate the ways in which this system may have changed over time and the ways in which the populations of these two prisons compared to their social environment. In its effort to classify people, the Bertillon system made use of a shifting schema of categories that cluster around the modern concept of race, such as "complexion," "nationality," "descent," "nativity," and "lineage." For this reason, it allows a window onto the historical development of the components from which modern ideas of race and ethnic identity have been assembled, and should shed light on how state power interacted with subject and marginalized populations in the late nineteenth and early twentieth centuries.

3 DOCUMENT ANALYSIS ON SCANNED BERTILLON CARDS

3.1 Data Pre-processing

The Bertillon cards used in this project were mostly scanned and stored in RAW image formats like Canon Raw format (CR2) and Olympus Raw Format (ORF). These were the outputs of the scanning devices in its original high resolution. Though the RAW image files contain more information, they are less convenient to use during image processing steps. The first step of preprocessing involved converting these images into a more usable format PNG. Though these images were taken under uniform lighting conditions, the brightness and contrast of the images had to be adjusted for better visibility of ink pixels. The original cards are at least 100 years old and because of the way they were stored, they have bent under their own weight, resulting in scanned images that were warped. The warped images were dewarped for getting straighter lines and text. All the preprocessing steps were carried out using ImageMagick tools [2].

3.2 Decimal Number Recognition

3.2.1 Overview. Front or recto side of a Bertillon card contains anthropometric data of the prisoner in two tabular sections (top and bottom), and frontal and profile view photographs of the prisoner. The top tabular section mainly contains quantitative bodily measurements like height, head width and length, etc., while the bottom section mainly contains descriptive details of the head like complexion, hair texture and color, build, ear and nose features, etc., (see Fig. 1). On the recto side of cards, for the work described in this paper, we concentrated on extracting handwritten decimal numbers from the top tabular section.

3.2.2 Card-type metadata files. As mentioned earlier, the card forms changed during the years, and about seven card forms were identified in the dataset. We created two kinds of metadata files, namely template and details files, for each of the card types that mainly describes the organization of the cards like position of horizontal and vertical solid lines, list of columns that need to be processed, their names, etc. These files were used for doing template matching and field matching. We also plan to include measurement units used in specific fields in these metadata files.

3.2.3 Segmentation. This step mainly involves segmentation of the scanned images into coarse region of interests and then later into finer image regions. For the recto side, this means identifying the top region of the card which contains tabular format details, then within that region, segmenting the sub-image into individual table cell images. These steps are briefly described below. The rotation correction and column image segmentation steps are based on an earlier work on census form processing [5]. Fig. 3 shows the output of various intermediate steps during segmentation.

Top tabular region extraction. As a first step of segmentation, we extracted the top tabular region of recto side cards. We observed that the height of the top region comes within a specific percentage range of the total card height and this is more or less uniform within a card type. Some amount of error margin was left to accommodate the cases wherein this relationship was not followed. Using this information, we cropped the top region images.

Rotation correction. The scanned card images were at times rotated because of slight misalignments and it had to be corrected to reduce that effect. The rotation angle was calculated by first identifying horizontal lines using Hough Transform, next calculating the angle of each line with respect to the horizontal axis, and then generating a histogram of these angles. The mean of the angles contained in the largest bin indicates the angle at which most lines were aligned. The top region image was then rotated in the opposite direction using this calculated rotation angle.

Column image segmentation. Through this step, we were able to segment column images from the top region image. We used concentration of ink pixels in the vertical and horizontal directions to roughly identify vertical and horizontal lines in the image. We then searched through the space of 2D rigid transformations, limited to scaling and translation, that best aligns with the lines found in the image to the ones in the template file. After the template matching step, individual column images were extracted (see Fig. 3).

Cell image extraction. After the column images were extracted, we filtered out only those column images that need further processing. The *details* file for a card type contains information about a card belonging to that card type like the total number of columns in the top region, columns that contain decimal numbers and hence which need further processing, the name and description of all the measurements in each of the columns etc. Based on the information in this details file, we sent relevant column images to be processed by the cell image extraction step. This step splits individual column images into their constituent cell images consisting of only one measurement value. We first performed morphological opening on each of the column images to remove spurious ink pixels. Later we did structural analysis on each column image to identify contour regions and after filtering out contours with area below a certain threshold, we found bounding rectangles around each of these contours. These rectangles were then grouped using K-means clustering technique wherein the number of clusters were set as the number of fields that need to be parsed in that column image. The rectangles within each class were then sorted in the increasing order of the Y coordinate values of their origin (top-left corner) pixel. Using this sorted list, we found a bounding rectangle for the entire class. Finally, the first and last horizontal line in this class

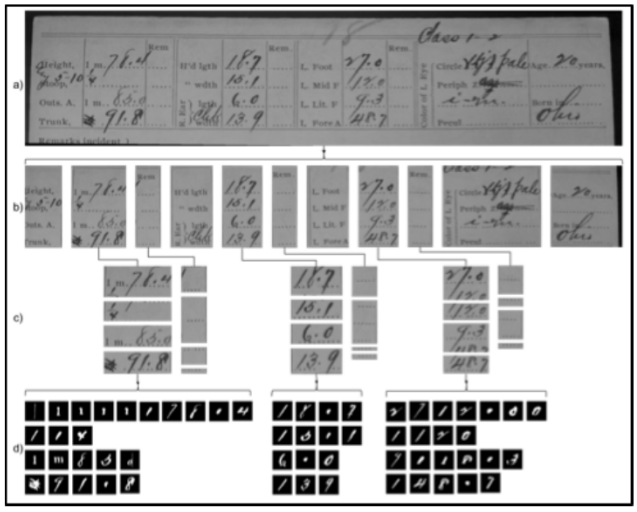

Figure 3: This figure shows the result of various steps in the segmentation process of decimal number recognition. Connectors, arrows and curly braces are used to show the relationships between different sub-images. a) Cropped and rotation corrected top region of a scanned card containing anthropometrical measurements. b) Column images generated after the column image segmentation step. c) Cell images extracted from column images that need further processing. d) Digit images isolated from cell images.

was used to extract a cell image. We found this approach to be better than identifying gaps between rows of numbers using vertical ink profiles and finding dips in ink concentration. The current approach enabled us to get around the cases when digits from one cell extends to a cell above or below it.

3.2.4 Digit image extraction. Once we have individual cell images, the next step is to isolate digit images that contain only digits or decimal points from each of the cell images. The approach here is pretty much similar to what was done in the previous step, except that the threshold values differ, and we also take care not to discard off decimal points by setting proper threshold values.

3.2.5 Digit classification and number recognition. After extracting digit images (including images containing decimal points), the generated images were sent to the digit classification module. Since there are many digits that need to be recognized in each card, we used an approach that gives competitive accuracy with fast recognition time [9]. This work employs Spatial Pyramid Histogram of Oriented Gradients (SP-HOG) feature vectors on Support Vector Machines to classify digits. The model used for classification was trained on the well-known MNIST database [8].

We used a heuristic approach to detect decimal points within a decimal number. On an average, the number of pixels or area needed to represent a decimal point and other dots were much less when compared with the digits from 0 to 9. Empirically we found this value to be 300 pixels in our dataset. Also, in our dataset decimal points are always preceded by at least one digit (e.g. '15.1'and not

something like '.15'). We combined these two pieces of information for detecting the position of decimal point within a decimal number.

3.2.6 Challenges. General challenges faced while processing the front side of the cards were predominantly variations in data organization between different cards and variations and imperfections within the cards. To address the former issue, we tried to identify different card types and created card template and details files for each of the card types, which documented these variations to be later used by processing programs. Since the processing program by itself is not able to identify these variations, we are adding more card types as we are discovering them.

Variations within individual cards were related to author's handwriting and style (e.g. in many cards we found that digit '2'was written in a slanted fashion and it highly resembled the digit 7), overwriting, written content extending beyond the allocated cell, ink smudges, torn cards, etc. We used morphological transformations to reduce the spurious ink pixels wherever possible, but we have come across situations where these smudges were big enough to be mistaken as numbers by the algorithm. To alleviate the effect of handwritten numbers extending beyond their allocated cells, we segmented individual digits along their border and placed them in a blank image before performing digit recognition. When we had just used bounding rectangles around these digits, there were more cases where these bounding rectangles contained unwanted ink pixels from neighboring numbers.

Finally, the presence of decimal points in the cards added more complexity to the algorithm. Individual rows of data are separated using dotted lines in these cards. Care was taken to make sure that only the dots in the dotted lines were removed during morphological operations and not the decimal point itself. Parameters used in this step were determined after processing a small sample of cards for many iterations.

3.3 Word Recognition

3.3.1 Overview. Performing OCR recognition for off-line, unconstrained, handwritten cursive text is notoriously difficult [11] where current specialized models might achieve up to 75% accuracy with large, clean datasets [6]. Thus, this task is somewhat exploratory as our goal was to get at least a measure of difficulty and at least some field transcribed. Upon inspection, the DESCENT field had a limited set of values. Therefore, we took the approach of trying to recognize whole words for this field as a start. In contrast to the front of the cards, the back of the card has no line boundaries that define 'cells', but instead dotted lines and field words, such as 'DESCENT .....', which together provide a general area for the writer to fill in. However, often the dots are covered by writing or smeared. Thus, the procedure details described here are similar to the front but also more tailored to the demands here.

3.3.2 Preprocessing. Preprocessing steps for the back side of the cards were overall similar to the front side. Card edges were found by finding edge contours. Due to some variance in brightness from the camera, it was found that a search from the middle of the image to the sides were better and more robust at picking out card edges. Templates were created for card layouts and details of the DESCENT field placement was created, however, in this case the

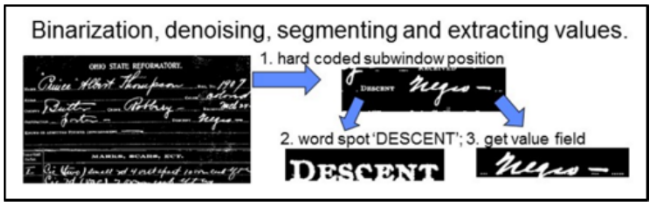

Figure 4: The segmentation and extracting process of handwritten word recognition.

details gave the general location of 'DESCENT'. Most card years there was only one template, however, some years, the templates varied from month to month, or even within a month. For these years, all segmentation was performed with two templates, and then the segmentation with the best word-spotting score of 'DESCENT' was chosen (see Fig. 4). Word-spotting is a technique in which a sliding rectangle was matched against a template image [7]. Once that best-matching rectangle position was found the cursive text was extracted as a window with dimension hard coded according to the template. In some cases, the image needed to be rotated 90 degrees, which also slightly affected the coordinates, even for the same template types. A final small rotation correction (less than 5 degrees) was applied based on pixel intensities. (This seemed to work better than a Hough transform in contrast to recto side of card because the dotted lines are not identified as distinctly as the solid lines.) These image segments were then binarized according to local mean threshold. Morphological operations were used as above to delete small objects of ink. The parameters for this deletion was found by trial and error to balance noise vs. signal. Occasionally, when the script is written faintly, letters become discontinuous and some of it gets deleted. The preprocessing and segmentation steps were run for groups of 2-3 years at a time. A time-consuming aspect of these steps was finding good parameters, lining up and confirming appropriate template parameters. These steps were coded in MATLAB and run on the Pittsburgh Supercomputing Center's Bridges supercomputer, where about 1000 cards take about 16-20 hours on one compute node.

3.3.3 Segmentation and Denoising Challenges. After the cursive text was mostly segmented and cleaned, additional preprocessing functions were added to handle specific noise elements. In fact, the varied sources of typical handwriting noise and the specific kinds of noise for these cards, are the main factors that make handwriting recognition difficult (see Fig. 5). A list includes the following: different writers, same writer sometimes using small vs capital letters, misspellings, misplaced or exaggerated features (i.e. dots for an 'i', or dash for 't'), text written over dotted line boundary, flourished writing that falls far outside expected boundaries, using abbreviations, extra dash after shorter words, and so on. As the recognition process proceeded, and as these noise elements were identified, the following were included in the final denoising process (see Fig. 6), along with caveats: using Hough transforms to identify dotted lines (but avoid deleting small lines that are part of script), remove marks from line above that hang partway down (but check that it is not connected to what could be an oversized first letter), remove dash-like extraneous marks (but check that they are elliptical and not

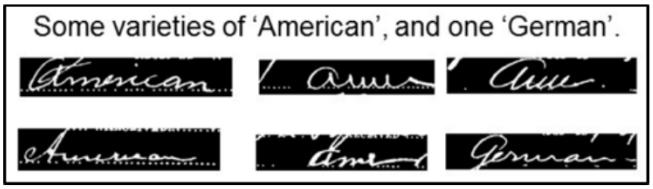

Figure 5: Examples of variance in handwriting. Notice different styles of 'A' and that the 'G' in 'German' is more like some of the 'A's. Also notice the hanging marks sometimes intermingle, as in the upper-left most case.

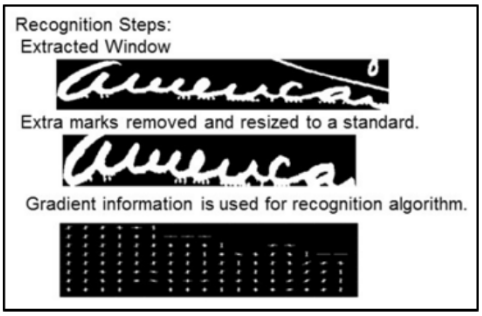

Figure 6: Depiction of final denoising and coding into features that are used as input for whole word recognition.

connected to larger components), remove empty space at left or right side of image, remove any left-over dots (but check that they are circular and not too large).

3.3.4 Word Recognition. The DESCENT field consisted primarily of a small set of values, namely, 'American', 'English', 'French', 'German', and 'Irish', along with numerous less frequent other cases. The approach taken was to treat whole words as the input pattern to learn. Using the segmentation process, with some extra manual-segmentation and manual-denoising, some clean training exemplars of 14 values were created from among the first years of data. Several classification techniques were tested, including Nearest Neighbor (NN) approaches (using both Euclidean based distance and dynamic edit distances, and 1-d profiles, i.e. column sums, of images); Naive Bayes (NB) using topological features (i.e. holes in the image); and Support Vector Machines (SVM) over edge-type information (histogram of gradients). After testing several parameter searches for each procedure SVM was found to work best when using DESCENT value images resized to 100x210 pixels. Using leave one out validation on the training set we got a best performance of 91% correct for SVM, 67% for NN with dynamic edit distance, 43% for NB. We tried combining all methods with a meta learner based on Random Forests but could not improve over the SVM scores. After creating a training set from the early years, the recognition process was run on following years. The process runs quickly, in large part because the final binary images of cursive script are small. After running the recognition process, the predictions were reviewed and corrections made. Also, as writers changed and certain values become more/less common, more training examples were added. For example, some writers started writing 'Amer' for 'American', so several cases were added to the training set; 'Austrian' started appearing more frequently in some later years, and some exemplars were added as well.

3.3.5 Ongoing Tasks for Recognition. The word recognition process is semi-automatic operating on the whole word with many tradeoffs worth considering. One is that it might be easier to just transcribe fields by hand, but that would be increasingly unwieldy as the number of images increase. Another approach is to use Mechanical Turk, online workers, but that would be increasingly expensive. Finally, one could also consider bootstrapping the DESCENT field transcriptions for training other fields. These will also be explored and discussed in the future.

4 RESULTS AND DISCUSSION

4.1 Word Recognition Results

The results for word recognition of DESCENT values demonstrate the challenges. After running the word recognition process, correct values were entered by hand, in about 1-2 hours depending on how well the prediction worked. That process is monotonous, but relatively quick except that the handwriting is sometimes difficult to read, even to this well-trained human. In fact, as is well known in psychology, letter recognition is highly dependent on semantic context, which is why OCR for such casual cursive script is difficult.

As preprocessing and segmentation was performed in groups of 2-3 years at a time (about 600-1000 cards per group), the word recognition was also run. Prediction results for the first groups gave accuracy between 60-70%. As more groups were processed more training exemplars were added depending on the handwriting styles, or new ways of writing, but performance also dropped to 45-55% for the later groups. Our final error rates are based on tests using the nearly full set of 6316 card backsides from the Ohio State Reformatory. In the end, there were 343 training exemplars created for 17 possible values. Most of these were manually cleaned, but sometimes hanging marks or dotted line portions were left in if it seemed minor. Of the 6316 cards, there were 383 cases of values that were not readable, and 196 that had poor DESCENT field segmentation, and 113 that had both. That means 383-113=270 cases were possibly well segmented but blank, poorly written, too faint, or contained non-descent values, such as the words 'don't know'. Also, 196-113=83 cases were poorly segmented (dynamic edit distance > 15), but still contained something readable. There were 702 cases that were not in the training set; 355 of which were double words, e.g. Scotch-Irish, which were left out to intuitively avoid making prediction too hard, and 702-355=357 cases that were relatively rare, e.g. 'Indian' or 'Bulgarian'. Future work on this dataset could look for double word marks, such as a dash, ampersand, or plus sign that often indicates two values combined. Overall there were about 150 unique categories of the DESCENT field values (e.g. where 'Amer' and 'American' are the same, and including double values such as Russian-Polish). Out of the remaining 6316-383-702=4888 predictable cases the recognition model produced a correct prediction 2731 times for about 0.56 accuracy. If we limit predictions to cases that had good segmentation (e.g. dynamic edit distance < 10) we get 2428 correct out of 5082 cases for an accuracy of 0.48. In other

words, good segmentation does not guarantee good predictions of the value. If we limit predictions to the cases where the SVM prediction score is high (e.g. > -0.005) we get 764 correct out of 1272 for an accuracy of 0.60. Thus, for a larger set, a possible strategy is to take the results for only strong predictions. However, one must be aware that low-frequency cases are not well predicted and likely will be missing from these results.

4.2 Decimal Number Recognition Results

We conducted pilot runs of the decimal number recognition extractor program on a subset of cards with ground truth labels and we are reporting the initial results based on these here[5]. The extractor program is integrated with Brown Dog and can be deployed to other HPC environments as well [10]. The results reported for the decimal number recognition step was obtained by running the extractor program on 206 cards from Ohio State Reformatory belonging to two different card types (125 cards belonging to Type1OSR1890s and 79 cards belonging to Type2OSR1900s). No specific criteria were used to choose these cards or card types. There were 24 fields that needed processing for each of the Type1OSR1890s cards, while there were 12 fields that needed processing for each of the Type2OSR1900s cards. Hence, a total of 3948 field values (12 x 79 + 24 x 125) were processed in the initial run. Some of these fields were remarks fields where the writer added any specific remarks about the measurement or if the original measurement was updated. Less number of those remarks fields are usually filled when compared with regular fields and they have been omitted in the current analysis for statistical reasons along with "English Height" or "Stoop" field, which contain mixed fraction instead of decimal numbers. We first report the percentage of cards in which the decimal point was detected (see Table 1). When multiple decimal points are detected within a field (because of other dots or imperfections), the algorithm recognizes more than one decimal number and when no decimal point is detected, it recognizes a number that won't generally make sense for that field of measurement (e.g. 1885 for height in centimeters wherein the actual number recognized should have been 188.5). Only those fields in which a decimal point was detected were considered for further analysis. We used mean absolute percentage error metric (MAPE) to quantify our results, since this metric will reflect how much each of the predicted value varies from its corresponding ground truth value. In other words, lower the MAPE value, better the recognition accuracy. The results are show in Table 2.

From Table 1, we can observe that the decimal point detection needs to be improved and it varies between fields and card types. This could be attributed to overwriting that is seen in many cards, dots in the dotted line present in the fields, and other spurious ink pixels. The accuracy of the decimal recognition step shown in Table 1 looks promising, but it should be considered that these results do not include the cards in which the decimal point wasn't detected in those specific fields and that these are results from a smaller set of cards used in pilot runs. This being a semi-automated system, the end user will still need to parse (may be programmatically) the results that are obtained from the system, but we believe that it is still better than manual labelling a huge number of cards.

[5] An extractor is a service that is part of Clowder, which is an open source research data management software. Please visit https://clowder.ncsa.illinois.edu/ for more details.

Table 1: Percentage of Cards with Decimal Point Detected

Field Name	Type1OSR1890s (%)	Type2OSR1900s (%)
Height	68.35	56.00
Outstretched arms	68.35	58.40
Trunk	49.36	40.15
Head length	56.96	35.20
Head width	72.15	51.2
Right ear length	60.75	46.4
Right ear width[a]	73.41	51.2
Left foot	26.58	13.6
Left middle finger	79.74	52.0
Left little finger	65.82	59.2
Left forearm	59.49	37.6

[a] Cheek width in case of Type2OSR1900s cards

Table 2: Mean Absolute Percentage Recognition Error for Different Fields

Field Name	Type1OSR1890s (MAPE)	Type2OSR1900s (MAPE)
Height	20.65	29.03
Outstretched arms	17.99	25.27
Trunk	7.52	10.29
Head length	17.52	10.56
Head width	12.64	15.93
Right ear length	12.08	26.31
Right ear width[a]	21.87	19.13
Left foot	17.43	17.23
Left middle finger	2.52	20.45
Left little finger	11.55	31.07
Left forearm	9.75	20.07

[a] Cheek width in case of Type2OSR1900s cards

4.3 Digital Humanities Results

In terms of results for the digital humanities, we noticed that it took the technologists longer than they had anticipated to overcome the variance presented by the cards, even though the cards look relatively systematic from a human point-of-view. This finding was important to the historical study of these objects because it drew attention to the actual variation happening within the Bertillon System over a relatively short period of time. On the other hand, this reality demanded that the technologists frequently tweak their workflows for each slightly different variation in card type. In the end, it is not clear how long it would take to get a trustworthy (and this is a critical word for historians) transcription of the cards through computational means. Historical data is already rife with potential error, and adding further computational error on top of this is a tricky proposition for humanists. Once the training and tweaking of the system is complete, the process of transcription would be quite short, of course, given the computing power at

the team's disposal, and it is possible that the amount of human error-correction needed by that time would be minimal.

5 CONCLUSIONS

To conclude, we have described the *Decomposing Bodies* project, its data, and research goals and how this work will shed light on some of the unanswered questions surrounding Bertillonnage as it was practiced in the United States and beyond. We have also proposed an end-to-end system for analyzing these cards in a semi-automated fashion, and have shared the current results from using the proposed system for automated transcription of Bertillon cards. Though far from perfect, we believe that the proposed system and its future enhancements will ease the burden on the researchers from doing tedious manual transcription when working on these cards.

This collaborative project has also demonstrated the importance of not only recognizing, but also deeply considering, the differing needs and reward structures inherent to humanists and technologists when working together. We have found that transdisciplinary projects like *Decomposing Bodies* need to plan for the time it will take to support the different ways that team members are motivated and rewarded. They then also need to make efforts on a continuing basis to ensure that all of the collaborators continue to get what they need out of the work.

That said, the critical importance of collaborations such as these is clear. The amount of digitized material now available to historians is a hot topic in the humanities, as many fewer researchers are finding it necessary to go out into the archives in order to perform their primary-source research. However, because this is the case, there is also a greater push to find new, sometimes highly computational, methods for using this digitized material to its fullest extent. Before, historians could potentially make their mark by finding an obscure archive and mining it for information that simply was not known to the rest of the field. Now that the historical dataset is becoming more and more easily digitally accessible, this has been changing the approaches of a growing number of "digital historians," including those working on the *Decomposing Bodies* project. Collaborations like this one, which attempt to completely transcribe handwritten records and then to digitally manipulate those transcriptions, are pushing the boundaries of what constitutes "something historians do."

It is not likely that the goal of Digital History is to train historians to be cutting-edge technologists as well as groundbreaking historians. The two disciplines of history and computer science have incredibly different training structures and work within disparate reward matrices. However, doing research and creating new knowledge between and amongst the assumptions and requirements of these disciplines is the job of both digital humanists and technologists who work with humanities data. Because this is the case, the computationally-intensive work that has been presented here is not the sort of research that can be done by most historians without access to XSEDE/ECSS support. It is our hope that collaborations such as this one become more prevalent and accessible in future.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grant numbers ACI-1053575 (XSEDE) and ACI-1261582 (Brown Dog).

REFERENCES

[1] 2016. Data (after)Lives: The Persistence of Encoded Identity. (2016). Retrieved 2017-02-23 from https://uag.pitt.edu/Gallery/81/theme/1

[2] 2017. Convert, Edit, Or Compose Bitmap Images @ Imagemagick. (2017). Retrieved 2017-02-21 from https://www.imagemagick.org/.

[3] 2017. Decomposing Bodies. (2017). Retrieved 2017-02-22 from https://sites.haa.pitt.edu/db/

[4] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *OpenFace: A general-purpose face recognition library with mobile applications.* Technical Report. CMU-CS-16-118, CMU School of Computer Science.

[5] L. Diesendruck, L. Marini, R. Kooper, M. Kejriwal, and K. McHenry. 2012. A framework to access handwritten information within large digitized paper collections. In *2012 IEEE 8th International Conference on E-Science.* 1–10. https://doi.org/10.1109/eScience.2012.6404434

[6] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. 2009. A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 5 (May 2009), 855–868. https://doi.org/10.1109/TPAMI.2008.137

[7] A. Kovalchuk, L. Wolf, and N. Dershowitz. 2014. A Simple and Fast Word Spotting Method. In *2014 14th International Conference on Frontiers in Handwriting Recognition.* 3–8. https://doi.org/10.1109/ICFHR.2014.9

[8] Yann LeCun. 1998. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/* (1998).

[9] Subhransu Maji and Jitendra Malik. 2009. *Fast and Accurate Digit Classification.* Technical Report UCB/EECS-2009-159. EECS Department, University of California, Berkeley. http://www2.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-159.html

[10] Smruti Padhy, Jay Alameda, Rob Kooper, Rui Liu, Sandeep Puthanveetil Satheesan, Inna Zharnitsky, Gregory Jansen, Michael C. Dietze, Praveen Kumar, Jong Lee, Richard Marciano, Luigi Marini, Barbara Minsker, Chris Navarro, Marcus Slavenas, William Sullivan, and Kenton McHenry. 2016. An Architecture for Automatic Deployment of Brown Dog Services at Scale into Diverse Computing Infrastructures. In *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale (XSEDE16).* ACM, New York, NY, USA, Article 33, 8 pages. https://doi.org/10.1145/2949550.2949647

[11] R. Plamondon and S. N. Srihari. 2000. Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 1 (Jan 2000), 63–84. https://doi.org/10.1109/34.824821

[12] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr. 2014. XSEDE: Accelerating Scientific Discovery. *Computing in Science Engineering* 16, 5 (Sept 2014), 62–74. https://doi.org/10.1109/MCSE.2014.80