Marriage Group Project Proposal

Jade Martinez, Taylor Schermer, Joel Nash, Lethicia Calderon, Lyla Wortham

September 8th, 2023

1 Extract

Genealogical research is beneficial for many studies relating to migration, family health history knowledge discovery, and insight into demographics. However, due to the physical degradation of these historical documents and the substantial volume, the development of transcribing records to online data files/bases is tedious and time-consuming. Our goal is to reduce the human load on transcribing these documents and increase the availability of the data for genealogical research.

Historical documents and manuscripts, including but not limited to marriage documents, can be accessed through the online database of many genealogical websites. In Phase One of the project, we would like to utilize Convolutional Neural Networks (CNN) to extract and index historical marriage manuscripts from these databases into a spreadsheet. In Phase Two, we will take the isolated data from Phase One and transcribe it into text. We will be using a data set consisting of historical marriage documents from Caiazzo, Italy dated between 1875-1889.

2 Literature Review

Dissection and transcription of historical manuscripts have become an increasingly popular topic of study within the past decade. There are various projects that are composed of different techniques to accomplish the task.

The transcription of historical handwritten marriage documents is not a new challenge; many different methods have been produced to approach the project. One study used a category-based bi-gram language model with implemented Kneser-Ney back-off smoothing to derive a connection from categories and semantic information within the record. They paired this with "document image processing, line image feature extraction, and Hidden Markov Model (HMM) and language model training/decoding" to achieve handwritten text recognition from "a collection of Spanish marriage license books conserved at the Archives of the Cathedral of Barcelona" (1).

Another approach applied line segmentation by utilizing a skeletal graph of the background of the image. An optimal path between text lines was determined using a path-finding algorithm. The goal was to create baselines and ground truth for further analysis. (2).

Another study pursues a different approach to word spotting that does not involve the segmentation of a document to retrieve data by using a graph embedding representation (3). Since structural approaches used to be slower than statistical ones, this method speeds up the matching by first locating the zones of interest from the image by using a graph embedding representation (3). They propose a segmentation-free word-spotting method based on graph representations, They

achieved this by using a Fourier-based descriptor and an alternative approach to RANSAC called PUMA (3).

Similarly, heterogeneous CNN with deep knowledge training was proposed to minimize "unreliable confidence in handwritten character recognition (4)" which is a problem associated with CNN when samples are poor. The CNN was trained with positive samples (characters) and negative samples (non-characters) to create more accurate results in recognition (4).

A fellow study developed two deep learning models in a sequence that can recognize distorted document images using European Language dictionaries. The model was trained with two differing data sets, one for text detection and one for text recognition. This improvement on previous models was shown through more accurate results (5).

3 Resources

The resources we plan to utilize for data acquisition and overall guidance of the project are listed below:

1. Ancestry.com
2. Daniel Faircloth (Translator)

3.1 Tools

1. **You only look once (YOLO):** Determines and predicts bounding boxes using end-to-end neural network.
2. **Roboflow:** A platform designed to streamline the process of managing and annotating documents for object detection.
3. **Convolutional Neural Networks:** A machine and deep learning algorithm that is utilized for classification and computer vision tasks.
4. **OpenCV:** Open source computer vision library used to extract data in each historical document.

4 Process

4.1 Tentative Implementation Strategy

1. **Data Annotation:** Team members collectively handled preprocessing, labeling, acquisition, and categorization/annotation, ensuring equal participation and learning opportunities.
2. **Coding Approach:** Utilized accessible algorithms and tools to ensure a shared understanding among the team, facilitating collaboration, task handling, and debugging.
3. **Documentation and Writing:** Encouraged all team members to actively contribute to research documentation and writing, fostering a comprehensive project understanding among everyone involved.

4.2 Group Interaction Platform and Strategy

To foster effective teamwork, our approach involved leveraging collaborative tools such as the Google *Colaboratory* platform and the Jupyter Notebook environment. These platforms enabled seamless access to shared code, facilitating group collaboration across diverse tasks. Additionally, we prioritized clear communication to define and assign tasks efficiently among team members.

5 Data Acquisition and Annotation

The marriage documents that were used for data acquisition were collected from the public database, ancestry.com. The chosen dataset used was acquired from Caserta, Campania, Italy, Civil Registration Records, between the years of 1862 and 1932. The selected data was from the years 1875 - 1879 for optimal feature extraction and annotation based on a similar document structure. Minimal pre-processing was required because of the quality of the documents from ancestry.com. Registers and annual indexes were filtered out of the data set and not used. Only fully completed marital documents were used. At least 10 documents from each year were annotated for training the CNN to identify handwriting in the post-processing phase.

The team utilized a collaborative project within Roboflow to efficiently manage task progress, successfully completing the data annotation phase. Each image was annotated by identifying printed keywords in each historic document and separating each region into 20 classes. The different classes are defined as the following: 2, Avanti di me, Numero, addi, anni, di, e d, figlia d, figlio d, l, meridiane, milleocentosettanta, minuti, nata, nato in, ore, presentat, presente, presenti, and residente. The team went through all of the annotated documents to verify each class location and ensure syntax homogeneity among individual documents. Consistent annotations defined a path for the CNN model to identify patterns of keywords in each document.

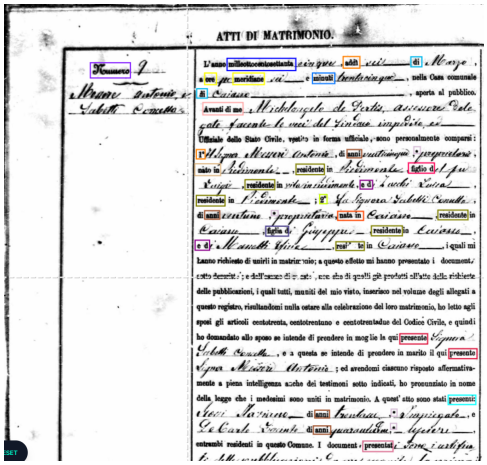

Fig 1. Annotations of marriage document from Roboflow highlighting keywords.

All team members contributed to the acquisition and annotation of data with the help of Daniel Faircloth for Italian to English translation. There was payment to obtain access to the public documents through ancestry.com. These are public documents. Therefore, there is an understanding that these documents connect to real people and will be treated with respect and appreciation.

6 Pre-Training and Testing

Separation of the marriage records into training, validation, and testing data sets was completed by the team in Roboflow. Once the annotations were verified, the annotated images were exported from the Roboflow and then trained on a YOLOv8 model.

The skeleton model of the CNN was adapted from the YOLOv8s.pt model, which is the small version of a pre-trained Region-based Convolutional Neural Network (RCNN) provided by YOLOv8. Using the annotated images, 27 images from our marriage record data set were implemented for our training data set. The validation data set used was composed of 17 images. To test the model, 46 images from various years were used.

7 Data Augmentation

Out of the original 90 images that were downloaded from ancestry.com, 80 documents were augmented to increase the size of the training data set in the YOLOv8 model. Augmentation was performed in Roboflow. The pre-processing performed in Roboflow consisted of image resizing of a 640x640 stretch. The augmentations performed to the original data set consisted of applied mosaic augmentation, a $\pm 5^{\circ}$ rotation, and a five-box cutout with a size of 25% each. An example of an augmented image can be seen in Figure 9.

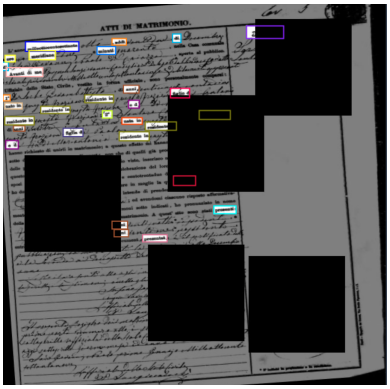

Fig 2. Original documents from ancestry.com after preprocessing and data augmentation processed were in Roboflow.

Data augmentation resulted in an improvement in the YOLOv8 model accuracy. The total size of the data set with the augmented data is 250 images. This allowed for a shift in the number of images in the training and validation sets in the YOLOv8 model. The model that the team trained next consisted of a training dataset with 80 original images and 160 augmented images. The remaining original images were used for the validation test, which now consists of 10 images.

The model was trained with 100 epochs, a batch size of 16, a learning rate of 0.01, and optimized with stochastic gradient descent. The results are displayed below.

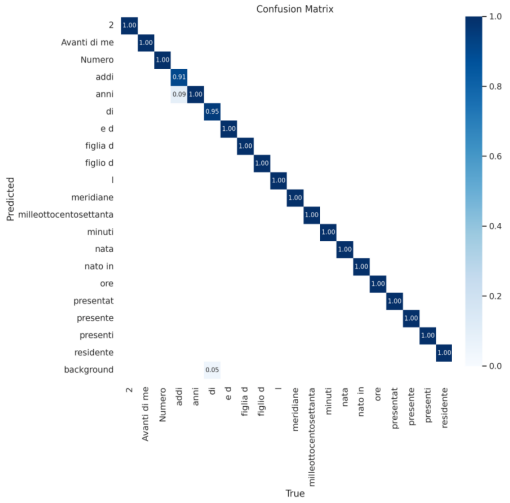

Fig 3. Confusion Matrix for the augmented marriage record data set after being trained in the YOLOv8s.pt model.

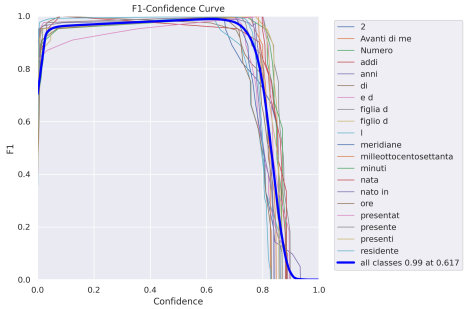

Fig 4. F1 Curve.

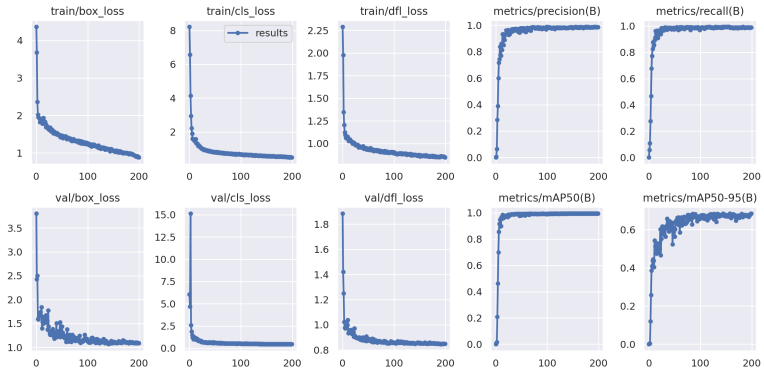

Fig 5. Loss, precision, and recall results from the training.

8 Data Extraction

Data extraction was performed using bounding boxes in OpenCV and used to crop the handwritten piece and transfer them to a virtual spreadsheet. Using the weights from the best-trained model, bounding boxes were used to identify the printed keywords in the documents. To extract the handwritten sections while still referencing their class, the bounding boxes around the printed keywords were shifted to bound the handwritten sections. The bounded handwritten portion was extracted and saved as a .jpg file to be stored under their correlated keyword. Headers were given to the keywords to be extracted, and placed at the top of the spreadsheet to easily identify the extracted portion of the document. The extracted .jpg and header files were vertically and horizontally concatenated to create a virtual spreadsheet. Lastly, exception handling was implemented. Since the keywords were not always located by the machine learning model, the size of the row would occasionally vary due to less cells.

9 Results

The team successfully built an algorithm which extracts important data from imported historical documents using a machine learning model from YOLOv8 and transforms the data into a virtual spreadsheet. This algorithm includes exception handling for when the data is not extracted from the document, and headers are placed on top of each column to define each extracted portion. The results of the algorithm tested on new historical documents are displayed in Figure 6.

Fig 6. Results of algorithm converting imported historical documents into a virtual spreadsheet.

References

[1] V. Romero and J. A. Sanchez, "Category-based language models for handwriting recognition of marriage license books," *2013 12th International Conference on Document Analysis and Recognition*, 2013.

[2] D. Fernandez-Mota, J. Almazan, N. Cirera, A. Fornes, and J. Llados, "Bh2m: The barcelona historical, handwritten marriages database," *2014 22nd International Conference on Pattern Recognition*, 2014.

[3] A. Hast and A. Fornes, "A segmentation-free handwritten word spotting approach by relaxed feature matching," *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 2016.

[4] S. Wang, L. Chen, L. Xu, W. Fan, J. Sun, and S. Naoi, "Deep knowledge training and heterogeneous cnn for handwritten chinese text recognition," *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2016.

[5] K. Mohsenzadegan, V. Tavakkoli, and K. Kyamakya, "A smart visual sensing concept involving deep learning for a robust optical character recognition under hard real-world conditions," *Sensors*, vol. 22, no. 16, p. 6025, 2022.