Extracting From Handwritten Historical Data from Registry Forms Using YOLOv8

Jade Martinez*, Lyla Wortham*, Dr. Nishatul Majid*, and Dr. Elisa H. Barney Smith†
*Physics and Engineering Department Fort Lewis College, Durango, Colorado, USA
{jumartinez, lpwortham, nmajid} @fortlewis.edu
†Wallenberg AI Autonomous Systems and Software Program
Luleå Technical University, Luleå, Sweden
elisa.barney@ltu.se

***Abstract*—Genealogical records can be a massive undertaking to search and index due to the sheer volume of historical documents that exist. Here we present an algorithm with high detection accuracy which extracts a set of target data points from a number of historical documents using an object detection network, referred to as YOLOv8, and transforms into a quickly glance-able virtual spreadsheet. This algorithm can be applied to numerous types of historical documents, which will contribute to quicker and less tedious genealogical research. We tested this approach on historical birth, marriage, and death documents. Here we present our result for marriage documents from Rio Grande do Sul, Barracão, Brazil dated between 1748-1998. In our study 110 total documents trained resulted in detection accuracy within two percent of the highest detection accuracy recorded which was 99.42%, which was a case with 80 raw training documents. Thus we were also able to evaluate the correlation between training dataset size and detection accuracy, which is useful in further studies to quantify the cost-benefit analysis of similar documents.**

I. Introduction

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 However, due to the physical degradation of these historical documents and the substantial volume, the development of transcribing records to online data files/bases is tedious and time-consuming [1]. Our goal is to reduce the human load when transcribing these documents to 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 Historical documents and manuscripts, including but not limited to marriage documents, can be accessed through the online database of many genealogical websites. Our team utilized a Convolutional Neural Network (CNN) and object detection algorithm to accurately extract and index historical marriage manuscripts from these databases into a spreadsheet. We utilized many different types of data, but this study focuses on historical marriage documents from Rio Grande do Sul, Barracão, Brazil dated between 1748-1998.

The robust algorithm presented here can provide a model for any type of historical document which provides accurate and more efficient means of any form of genealogical research. The reduction of time taken when sifting through historical documents is crucial when working to locate specific information. An accessible spreadsheet generated with user input data is less intimidating and tedious than repetitively funneling through an abundance of historical documents with key information being located only by the human eye. Especially for large data sets, the process described in this research shows promising efficiency improvements evaluation of lineage and general genealogical research.

As the use of CNN for genealogical research and other modern advancements continues to gain popularity, there is little information available which reflects the minimum amount of data needed to develop a CNN model with sufficient detection accuracy. To provide information regarding data set size, we also worked to develop a graph depicting the correlation between model detection accuracy and training data set size.

Object detection networks have proven to be useful in many different areas of research involving offline handwriting recognition (OHR). In offline handwriting recognition, features are generally extracted from the input data and matched to a sequence of labels (i.e. alphabet characters), typically using a neural network [2]. YOLOv8 (You Only Look Once) is is an object detection model that is fundamentally a CNN structure often utilized for its balance of speed and detection accuracy [3]. Both elements are crucial when processing keywords in a large amount of data, such as the historical datasets utilized in this project.

By sifting through historical documents and identifying information on each page, we were able to present essential information in a more accessible manner. A virtual spreadsheet was compiled by extracting the most relevant information from multiple historical documents using a CNN model and combining that information into uniform rows and columns which are easily processed by the human eye. English category labels were included in our compiled spreadsheet for ease in indexing desired data such as names of individuals, their spouses, birth dates, birthplaces, etc. The overall workflow process for this compilation project is documented in Figure 1.

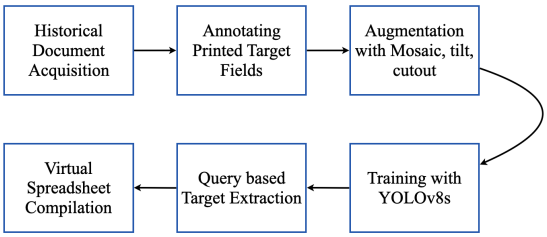

Figure 1. Block diagram of the workflow.

II. 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. Generally deep learning-based methods are applied to text recognition and detection tasks. One of the latest trends in text detection is to use CNN [4].

Efforts have been made in multiple studies to streamline the text recognition process when applying Optical Character Recognition (OCR) to historical documents. Typically a neural network generates bounding boxes on the scanned in the document according to chosen parameters, then an OCR is applied to these same boxes to analyze the contents [5]. There have been more recent efforts to meld the two processes. A more recent study combined the two separate processes, segmentation and recognition, into one process- a proposed Simple Predict and Align Network. This allowed for a more efficient training [6]. Many object detection algorithms are used to revolutionize computer vision and object detection. One project utilized YOLOv8 to predict the location of bounding boxes. This project was based on precision-recall metrics, handling multiple object categories, and defining a positive prediction using Intersection over Union. By tracking how the mean average precision changes over time with different guideline inputs of the model, the detection accuracy of the model can be compared and contrasted over time [7].

The transcription of historical handwritten marriage documents is not a new challenge; many different techniques have been used 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 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" [8]. 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. [9].

Some projects focused on the speed and amount of data taken to develop successful models. One method speeds up the matching by locating the zones of interest from the image by using a graph embedding representation [10]. Another approach included the proposal of a heterogeneous CNN with deep knowledge training to minimize "unreliable confidence in handwritten character recognition", which is a problem associated with CNN when samples are poor [11]. 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 [12]. Another study proposed geometric factors which were key to faster and better performance in their model [13].

Data annotation and information extraction through machine learning is utilized to find and classify relative information. One study utilized information extraction to build domain-specific search engines using titles, authors and research paper headers [14]. A water research study used data annotation techniques to locate and extract anomalies in environmental systems [15]. Another study employed a CNN to detect and recognize low quality licence plates through data annotation for traffic control benefits [16].

III. Data Acquisition and Annotation

The marriage documents that were used for data acquisition were collected from a Brazilian public dataset, dated between the years 1748-1998 [17]. The primary data used for CNN model training was selected between the years 1929-1948 for optimal feature extraction and annotation based on the samples having similiar structure. Minimal pre-processing was required because of the quality of the documents from FamilySearch.org. Registers and annual indexes were filtered out of the data set and not used. Only fully completed marital documents were used.

For each acquired sample, ground truth was defined by annotating the document in its original state. The samples were annotated by identifying the target field of printed keywords in the historical document and categorizing each field into one of 16 classes. For 36 instances of printed keywords in each sample, target fields for each instance were defined by recording the x and y values in each document. The target fields were mapped to their correlating class. The ground truth samples, or annotated original documents, were annotated and stored using a computer vision platform, Roboflow [18]. The printed keyword classes are defined and translated in Table 1.

Portuguese Keyword	English Translation
19	19
aos	on the
de	of
domiciliada em	(she) was living in
domicliado em	(he) was living in
e	and
em	in
filha de	daughter of
filho de	son of
juiz	judge
matrimonio de	matrimony of
nascida em	(she) was born in
nascido em	(he) was born in
profissao	profession
residente em	resident of
testemunhas	witnesses

Table 1. English translation of selected Portuguese keywords in Brazilian marriage documents.

Note that Portuguese characters were not included in the class labels when annotating. All annotated documents were carefully evaluated 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. A ground truth historical document from the Brazilian marriage dataset [17] is shown in Figure 2.

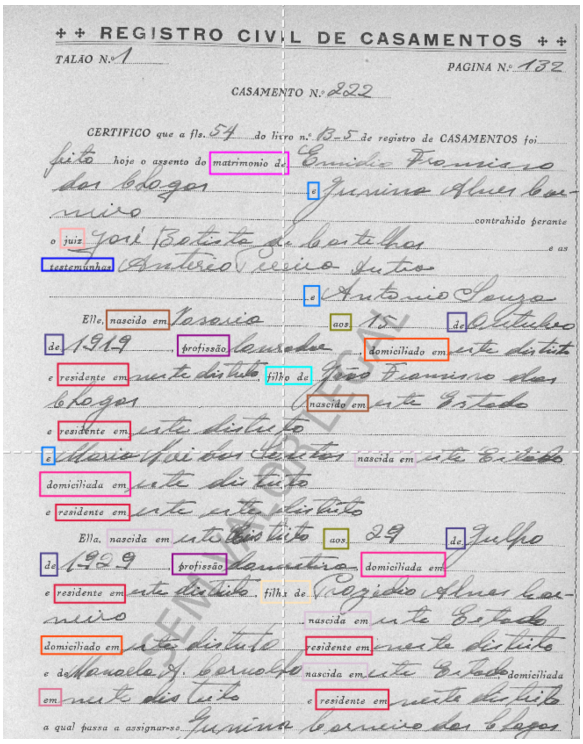

Figure 2. Brazilian marriage document with printed keywords annotated [17].

Completed annotations allowed for pre-processing applications to be applied to the data. Augmentations and post-processing the data increased the overall size of the data and allowed for easier implementation to a computer vision framework.

IV. Data Augmentation

Overfitting when training a CNN model occurs when the model begins to memorize the training samples, thus performing poorly on new data. The data used in this study was prone to overfitting because of the invariability between each sample. Data augmentation is a type of data manipulation commonly utilized for CNN model development. More data is added to the raw data by manipulating the sample images in various ways. Robust variations of data produce an improvement to the CNN model by reducing overfitting. The variations of the raw samples also allow for more accurate detection for more diverse data.

Out of the 100 raw samples that were downloaded from FamilySearch.org, 90 documents were augmented to increase the size of the training data set in the YOLOv8 model. The pre-processing functions executed consisted of image resizing of a 640x640 stretch. Using a 1:2 raw to augmented data ratio, 270 images from our marriage record dataset were implemented for our training dataset. The validation data set used was composed of 10 raw images. The following augmentation manipulations were applied to each of the raw samples in the dataset:

- Mosaic
- ±5° rotation
- Five-box cutout with a size of 25%

V. Training

The skeletal 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. We used the skeletal model as the base for out CNN. The samples in the data set were separated in a 90:10 training to validation split. Training samples were fed to the CNN and the detection accuracy of the developed model was tested with the validation dataset. CNN models were developed with a 0.1 learning rate. Detection accuracy was evaluated through precision metrics and accuracy measurements.

VI. Virtual Spreadsheet Compilation

Using the weights from the best trained CNN model, printed keywords were predicted in unseen documents. The fields predicted by the CNN model were used to define bounding boxes aligned with the target field for the most relevant printed keywords. Bounding boxes were manipulated to confine the handwritten portion relating to the printed keyword. The bounding boxes were then extracted and saved as files to later be compiled into a virtual spreadsheet. Data extraction and concatenation was performed using OpenCV, the open source computer vision software library [19]. Predictions with the confidence for each instance for printed keywords run through the YOLOv8 model in an unseen document are shown in Figure 3.

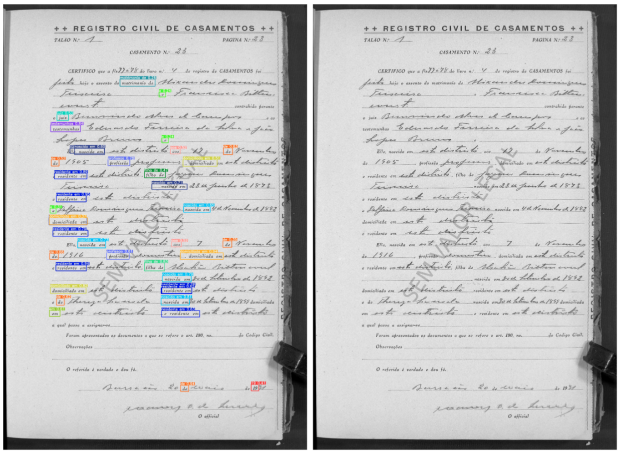

Figure 3. Brazilian marriage document with predicted printed keywords after being passed through the YOLOv8 model compared to the original marriage document.

In this study, the printed keywords selected were matrimonio de, de, e, filho de, filha de, nascido em, and nascida em. These keywords were used to define the both the husband and wife's names, birthdays, birthplaces, and parents. To extract the handwritten sections while still referencing their class, the bounding boxes around the selected printed keywords were shifted to bound the handwritten sections and extracted. The extracted files were saved as a .jpg file to be stored under their correlated keyword for uniform spreadsheet compilation.

Some keywords had multiple instances which were important for information extraction. In these cases, the regions of the predicted keyword were delimited, and the bounding box was shifted and defined based on the desired extraction of the handwritten portion. The extracted .jpg files were also regionally defined by the location of the correlated bounding boxes and named with compatible labels.

Headers were assigned to the extracted keywords 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. Since the keywords were not always located by the machine learning model, the size of the row would occasionally vary due to less cells. To avoid inconsistent row sizes, an error handling .jpg file was used in place of the document's bounding box if the keyword was not predicted by the model.

While our team extracted 13 categories, the virtual spreadsheet can also generate on requested fields of entry only. For example, if the user chooses to search for only the names of the women and their birthplace, only those categories will be shown for the input documents.

The data extraction and spreadsheet development algorithm was tested on 23 historical documents. These documents were new, meaning that have not been seen by the best YOLOv8 model. A virtual spreadsheet generated from unseen data is displayed in Figure 4.

Figure 4a.

Figure 4b.

Figure 4. Virtual spreadsheet with extracted information regarding the wife and the marriage date, using extracted handwritten data from 23 unseen Brazilian marriage documents.

VII. Results

We successfully built an algorithm which extracts important data from imported historical documents using a machine learning model in YOLOv8 and transforms the data into a virtual spreadsheet. This algorithm included exception handling for missed data predictions and defining spreadsheet headers. We were also able to evaluate the correlation between training data set size and detection accuracy, which proved to be useful in studies which include extensive data annotation and acquisition of a large amount of data.

A. Detection Accuracy and Data Extraction

While our data extraction algorithm worked well when shifting and extracting data, our CNN model was also very successful. The overall detection accuracy of the model trained with 200 epochs resulted in our highest detection accuracy, 99.625%. This model is referred to as our best model. This section will discuss the metrics, detection accuracy and precision of our best model.

A few metrics were used to evaluate the detection accuracy of the CNN model. We used the F1 Score, Precision-Recall (PR) Curve, and the confusion matrix to gain insight on the abilities, detection accuracy and precision of the model. The F1 Score is the metric that displays the mean of the precision and recall of the model. Precision is the CNN model's ability to avoid false positives, while Recall describes the models ability to detect instances of all classes. While the F1 score is a single metric at a specific threshold, the PR curve shows the trade off between precision and recall at different thresholds. Confusion matrices provide insight on the overall detection accuracy for each instance in a CNN model.

The F1 Curve showed an ideal mean of the precision and recall. The PR Curve of the model also showed an almost perfect output for each threshold.The Confusion matrix for the model showed perfect detection for almost fourteen of the predicted classes. The predictions for class “e” showed a 97% overall prediction accuracy. The predictions for class “em” showed a 91% prediction accuracy.

B. Impact of Data Set Size on Performance

While the use of CNN models to induce advancements in genealogical research can alleviate the extensive workload of researchers, the initial preparation of the CNN model can be very tedious. During the data acquisition phase, it is crucial to obtain data that is consistent and does not include registers or cover pages. Various types of historical documents tend to be prevalent in databases containing miscellaneous data. Filtering data can take the developer various amounts of time, depending on the type of data included in the data set. Filtering documents from large, miscellaneous databases can take upwards of an hour in some cases.

The amount of data needed to train the data is crucial to obtain sufficient detection accuracy. With less documents to obtain from the desired database, less data filtering and less data annotation is likely to reduce data preparation time. Correlation between the amount of data and YOLOv8 model detection accuracy was studied in this project through a model with the same raw Brazilian marriage documents used for best model, but different training criteria.

During the preparation phase for a CNN model used with an approach such as this one, the amount of data used to train the model is linked directly to the detection accuracy and training time. Evaluation of desired detection accuracy can be very advantageous in minimizing the time taken to prepare data for CNN model training.

C. Performance and Time Correlation

Depending on the type of data and annotation strategy, the annotation phase can also take extended period of time. For this particular project, the annotation strategy required annotation for all printed key words located before handwritten sections. Finding the ground truth by annotation for an individual document took between four to five minutes. A total of 100 unprocessed documents were annotated, 90 of which were filtered into an augmentation post-processing group and 10 into the validation data set. This process was relatively tedious. The total amount of time taken to annotate all data for the best model, not including annotation verification or data splitting, was approximately 6.5-8 hours. Each document contained 16 classes with 36 instances. Annotation with less instances in each document could take less time, depending on the amount of data being annotated.

The model with the best detection accuracy was trained with 200 epochs, a batch size of 32, a learning rate of 0.01, and optimized with stochastic gradient descent. The model was trained in the Google Colaboratory hosted Jupyter Notebook environment using the Python 3 A100 GPU runtime type.

Another model was developed using the same raw data. This model was used to compare the size of a data set with its detection accuracy and run time. The models used to evaluate the impact of size were trained in the Google Colaboratory hosted Jupyter Notebook environment using the Python3 A100 GPU runtime type. Every model was trained with 100 epochs, a batch size of 32, a learning rate of 0.01, and optimized with stochastic gradient descent. The results of the data set size and detection accuracy correlation testing are in Figure 5.

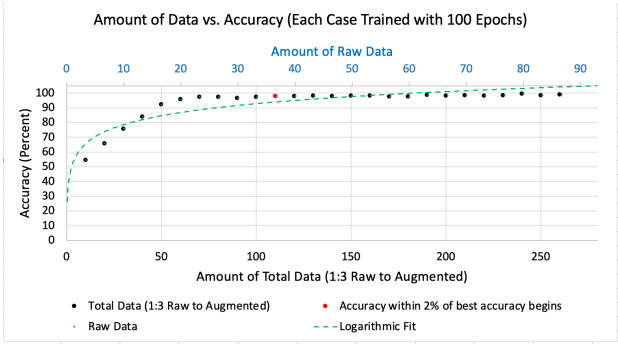

Figure 5. Correlation between YOLOv8 model detection accuracy and the amount of data for both raw data and augmented data.

Note that sufficient detection accuracy varies based on the intent of model developer and the purpose of the model. Figure 5 emphasizes the case at 110 total data (1:3 raw to augmented data) trained. This case was the first case within two percent of the highest detection accuracy. Each case following slowly improved until a 99.42% was reached at the case trained with 240 total training documents. The CNN model with the most data, 260 total images, reached a detection accuracy of 99.12%.

Another issue relating to time when training the model is the time taken to process all of the data. Less data ideal when aiming to reduce the amount of time necessary to train the model. For the best model, the training process took 37.2 minutes. Correlation between run time and the amount of data was also evaluated in this project. The cases trained using 100 epochs and used to evaluate the data in Figure 5 were also used to implemented run time analysis measurement, which resulted in a linear increase in time vs. data being trained.

VIII. Conclusion

We successfully built an algorithm which identifies and extracts user chosen categories from historical documents using a machine learning model in YOLOv8. The algorithm also compiled this data into a virtual spreadsheet that can be utilized for more efficient research. Transformation of a raw data set into a virtual spreadsheet allows for an improvement in batch processing of large data sets, quick and efficient data searches, and sorting and indexing of data sets. We were also able to evaluate the correlation between training data set size and detection accuracy. The data set size comparisons proved to be invaluable when it comes to time management for batch processing and acquisition of large amounts of data. Similar published research was not found to compare our approach

and results.

This process has shown promising potential benefits for future genealogy research. Extracting information from registries like this can take specialists and volunteers many hours. The necessity to change focus and find the position in the form where the desired information is located, copy it to a record, return focus to the document, and find the next field during the indexing adds to processing time. By extracting the data from the fields and pre-sorting it, the change of focus steps are not needed. If a researcher is looking for specific family names, this virtual spreadsheet can be quickly scanned to find the records of interest. Rather than sifting through historical documents by hand, an indexed spreadsheet allows the researcher to access quicker and more efficient results. The time taken for compilation of a spreadsheet with this algorithm is minuscule compared to the acquisition of key data points by hand. For both amateur and professional genealogy searches, efficiency when searching for results is ideal.

Digitizing historical documents has been a goal for decades across sectors, especially family history. Finding more efficient ways to do so is crucial work. In the future, we are planning to implement keyword spotting to this algorithm for ease in category lookup. We will also aim to streamline the data augmentation process. By utilizing our framework, we also plan to process different types of historical documentation for further research, ie tax records.

Acknowledgements

We thank Taylor Schermer, Joel Nash, Lethicia Calderon, and all members of the 2023 Computer Engineering Junior Design Team at Fort Lewis College for their contribution to this project, Dr. Cesar Da Silva for Portuguese to English translations, and Dr. Randy Wilson at FamilySearch for his assistance with data set permissions.

References

[1] D. R. Wilson, "Bidirectional source linking: Doing genealogy "once" and "for all," 2002. [Online]. Available: https://api.semanticscholar.org/CorpusID:15542129

[2] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., vol. 21. Curran Associates, Inc., 2008. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2008/file/66368270ffd51418ec58bd793f2d9b1b-Paper.pdf

[3] J. Deng, X. Xuan, W. Wang, Z. Li, H. Yao, and Z. Wang, "A review of research on object detection based on deep learning," *Journal of Physics: Conference Series*, vol. 1684, no. 1, p. 012028, nov 2020. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1684/1/012028

[4] R. Li, S. Chen, F. Zhao, and X. Qiu, "Text detection model for historical documents using CNN and MSER — IGI Global," *Journal of Database Management (JDM)*, 2023. [Online]. Available: https://www.igi-global.com/pdf.aspx?tid=322086&ptid=310090&ctid=4&oa=true&isxn=9781668478929

[5] B. H. Marti, UV., "The IAM-database: an English sentence database for offline handwriting recognition," *IJDAR*, vol. 5, pp. 39–46, 2002.

[6] D. Coquenet, C. Chatelain, and T. Paquet, "SPAN: a simple predict & align network for handwritten paragraph recognition," in *IEEE International Conference on Document Analysis and Recognition*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231942377

[7] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1680–1716, 2023. [Online]. Available: https://www.mdpi.com/2504-4990/5/4/83

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *AAAI Conference on Artificial Intelligence*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:208158250

[14] A. McCallum, K. Nigam, J. Rennie, and K. Seymore, "A machine learning approach to building domain-specific search engines," in *IJCAI*, vol. 99, 1999, pp. 662–667.

[15] S. Russo, M. D. Besmer, F. Blumensaat, D. Bouffard, A. Disch, F. Hammes, A. Hess, M. Lürig, B. Matthews, C. Minaudo *et al.*, "The value of human data annotation for machine learning based anomaly detection in environmental systems," *Water Research*, vol. 206, p. 117695, 2021.

[16] J. Špaňhel, J. Sochor, R. Juránek, A. Herout, L. Maršík, and P. Zemčík, "Holistic recognition of low quality license plates by CNN using track annotated data," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2017, pp. 1–6.

[17] FamilySearch, "Brasil, Rio Grande do Sul, Registros Diversos, 1748-1998, Barracão, Matrimônios, Talão 1, 1929-1948, images," https://familysearch.org/ark:/61903/3:1:33S7-81RD-S86?cc=1985805&wc=QZS2-VVQ%3A264195001%2C1588922507:22October2015.

[18] J. S. B Dwyer, J Nelson, "Roboflow." [Online]. Available: https://roboflow.com

[19] G. Xie and W. Lu, "Image Edge Detection Based On Opencv," *International Journal of Electronics and Electrical Engineering*, vol. 1, pp. 104–106, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:35744810