Crowdsourcing Image Datasets: An Examination of Ground-Truth in Labeling, Text Segmentation, & Sampling Bias

Matthew I. Swindall

Middle Tennessee State University
mis2n@mtmail.mtsu.edu

Abstract

Character recognition and text extraction methods remain impractical for ancient, handwritten manuscripts. Within the fields of digital humanities and papyrology, no artificial intelligence tools currently exist that can effectively aid the scarce pool of experts in the rapid transcription of vast volumes of existing documents. In my previous and proposed work, I seek to address the underlying challenges impeding development of such tools. A pipeline for the creation of crowdsourced image datasets is proposed that will mitigate the effects of sampling bias and ground-truth uncertainty inherent in the crowdsourcing process. Such datasets can then be used to train models which can be employed by interactive agents to accelerate document transcription in cooperation with trained human experts. The pipeline consists of volunteer annotations via a web-based interface, consensus labels and locations derived from these annotations, synthetic augmentation of the data to reduce sampling bias, a misclassification classifier that can reduce uncertainty in ground-truth labeling, and a practical method for segmenting text in damaged manuscripts. The creation of such a pipeline will require multiple challenges with no current solution to be overcome. Additionally I hope to show that such an approach can be generalized to other datasets. My proposed series of experiments are designed to probe the viability of such a system and encourage others to contribute to further explore possible approaches.

Keywords— Crowdsourcing, Human-Computer Interaction, Digital Humanities

1 Motivation

While character and text recognition are mature fields, most models are trained on carefully curated or manufactured, custom datasets. Ancient documents however are a vastly different quality of data source. Manuscripts are often incomplete and damaged, exhibiting holes, tears, and faded or smudged text. These documents are challenging if not impossible for state-of-the-art systems to decipher. It is of immense importance to the study of ancient texts, and the history they contain, to create systems capable of rapidly transcribing extensively damaged manuscripts. There exist vast collections of un-transcribed documents from the ancient world, yet very few experts qualified to transcribe them. Crowdsourcing annotations of ancient documents is an important step in developing such systems, but such efforts introduce a great deal of uncertainty in ground-truth labeling. Volunteer annotators regularly make incorrect judgments and are often in disagreement with each other. Qualifying and quantifying the trustworthiness of volunteer annotations is a necessary next step toward practical use of crowdsourced datasets. In addition to ground-truth uncertainty, crowdsourced datasets contain additional noise including sampling bias. Some characters are less common than others in a given language, or in a given collection of documents. For the rarest characters, class imbalance ensures models yield low accuracy for such samples. Automatic text extraction is yet another challenge for ancient documents, as text recognition engines are easily confused by the damage in these manuscripts. Solving the problems of ground-truth uncertainty, sampling bias, and text extraction for ancient manuscripts would make it possible to design interactive agents capable of assisting experts in transcribing these documents and dramatically reduce the time necessary to transcribe entire collections.

2 Background & Related Work

The Ancient Lives Project was a web-based crowdsourcing initiative that allowed volunteers to transcribe digital images of ancient papyrus fragments from 2011 until 2018 (Williams et al. 2014) in coordination with the Zooniverse (Simpson, Page, and De Roure 2014). This Project resulted in millions of annotations from images of papyrus fragments. These manuscripts consist of handwriting on papyrus, in various states of preservation, excavated from rubbish dumps near the ancient Egyptian city of Oxyrhynchus (Bowman et al. 2007). The Oxyrhynchus Papryi collection is estimated to contain roughly 500,000 fragments, of which only a tiny fraction have been transcribed since their discovery in the late 1800's. A consensus algorithm was applied to these annotations to produce approximate locations of individual characters within the document images, as well as consensus labels for each identified character.

The data were then utilized to created the Ancient Live Dataset (Swindall et al. 2021). In order not to reveal manuscript data still under papyrological research, two versions of the dataset were created. AL-PUB is derived from previously published material in The Oxyrhynchus Papyri Series while AL-ALL is derived from both published and unpublished manuscripts. As such, only the AL-PUB dataset was made available to the public. Each image in both dataset versions *ideally* contains one tightly cropped Greek character, stored in JPEG format, and can be of a range of several resolutions. All 24 characters in the standard Greek alphabet are represented in the dataset, as shown in Figure 1. AL-PUB contains 195,683 character images while AL-ALL contains 399,421 images.

The datasets both exhibit noise in the form of ground-truth uncertainty and extreme sampling bias. The consensus algorithm utilizes majority vote for class labeling. In many instances there is a great deal of disagreement between volunteer annotators which yields imperfect, but effective results for our cropping algorithm. Sample sizes are widely varied, as demonstrated in Table 1, with

Figure 1: Examples of Character Images from the AL-PUB Dataset

Figure 2: Ambiguous Tau Instances. From left to right: Incorrectly rotated, possibly a Pi, multiple characters present.

the largest sample in AL-ALL containing 46,344 images while the smallest contains only 62. Additionally a small percentage of images may be completely devoid of any characters, contain partial characters, or contain multiple characters. Such outliers may have resulted from improperly rotated images or incorrectly recorded location coordinates. Manually identifying and removing such images has proven challenging as there are many ambiguous instances. Figure 2 shows examples of ambiguous images labeled as Tau.

To mitigate the effects of class imbalance in AL-ALL, two of the smallest samples were augmented by adding synthetic instances generated via StyleGAN2 in (Swindall et al. 2022). This work illustrated an interesting and not yet understood relationship between per-class accuracy and overall model accuracy. While the overall training and validation accuracy were virtually identical for models trained with the original data and with the synthetically augmented data, per-character accuracy for the augmented samples increased by 8% and 12%. This likely suggests that augmenting one sample may negatively effect accuracy for other samples. Further work is necessary to better understand this relationship.

To mitigate the effects of ground-truth uncertainty in AL-ALL, my team is utilizing an ensemble approach to better understand the uncertainty. We apply stacked generalization consisting of nearly identical ResNets: one utilizing cross-entropy (CXE) and the other Kullback-Liebler Divergence (KLD). The CXE network uses standard labeling drawn from the crowdsourced consensus. In contrast, the KLD network uses probabilistic labeling for each image derived from the distribution of crowdsourced annotations. We refer to this labeling as the Human Softmax (HSM) distribution which is similar to an approach taken in (Peterson et al. 2019). For our ensemble model, we apply a k-nearest neighbors model to the outputs of the CXE and KLD networks. Individually, the ResNet models have approximately 93% accuracy, while the ensemble model achieves an accuracy of $>$95%. We also perform an analysis of the Shannon entropy of the various models' output distributions to measure classification uncertainty. The results suggest that the Shannon Entropy of prediction probability distributions may be a good way to quantify the trustworthiness of each label. The hope is that quantifying label trustworthiness in this way will allow for the removal of undesirable images from the dataset. This work is currently ongoing.

Character	Count	Character	Count
Alpha (A, α)	42,546	**Nu** (N, ν)	44,910
Beta (B, β)	2,534	**Xi** (Ξ, ξ)	1,201
Gamma (Γ, γ)	6,907	**Omicron** (O, o)	46,344
Delta (Δ, δ)	11,717	**Pi** (Π, π)	17,114
Epsilon (E, ϵ)	31,584	**Rho** (P, ρ)	20,450
Zeta (Z, ζ)	1,425	**Sigma** (Σ, σ)	62
Eta (H, η)	15,064	**Tau** (T, τ)	32,045
Theta (Θ, θ)	7,575	**Upsilon** (Y, υ)	15,762
Iota (I, ι)	25,595	**Phi** (Φ, ϕ)	6,063
Kappa (K, κ)	17,937	**Chi** (X, χ)	9,156
Lambda (Λ, λ)	13,253	**Psi** (Ψ, ψ)	904
Mu (M, μ)	13,227	**Omega** (Ω, ω)	16,046

Table 1: Counts for each letter in the Ancient Lives dataset.

3 Proposed Research

In my proposed research I seek to address four main challenges:

1. Quantify the trustworthiness of volunteer annotations and model prediction probabilities
2. Improve text extraction for ancient, handwritten documents
3. Reduce the effect of sampling bias on crowdsourced datasets
4. Develop a pipeline for the crowdsourcing of image datasets

To understand the trustworthiness of volunteer annotations and model prediction probabilities, I propose the creation of a misclassification classifier. I envision this as a model that, given prediction probabilities from a model or human annotations for a given image, the classifier will quantify the likelihood that the ground-truth is correct.

In an effort to improve text extraction for ancient handwritten documents I propose an exploration of automated text line segmentation and spectral clustering using text locations from the Ancient Lives Project as techniques for segmenting text in such documents, followed by deep learning approaches for identifying individual characters. My hope is that combining these methods will allow for automatic and rapid annotation of damaged, ancient documents such as those in the Oxyrhynchus Papyri collection.

Generative networks are likely the best approach for reducing sampling bias in AL-ALL. A combination of generative GANs and Neural Style Transfer GANs may be an effective approach for synthetically augmenting extremely small samples in crowdsourced datasets, resulting in higher per-character accuracies for classification models.

Upon solving the first three challenges, I believe an accurate and practical pipeline may be created for crowdsourcing image datasets based on ancient, handwritten documents, that can be utilized to partially automate the transcription of additional documents in cooperation with trained experts in the field. The ideal form for such a system is an intelligent AI agent that works with the expert to accelerate the transcription process.

4 Proposed Experiments

4.1 Misclassification Classifier

This experiment will make use of work currently in progress, which utilizes the Shannon Entropy of an ensemble model, along with the Human SoftMax probability (HSM), derived from the original volunteer annotations, to quantify the trustworthiness of model predictions and the ground-truth labels for the AL-ALL dataset. The goal

is to remove images from the dataset which the entropy suggests is likely to be misclassified by the model, by the annotators, or by both. The expectation is that removing misclassified images may lead to a more accurate classification model. A counter-hypothesis is that removing the "bad" data may actually result in decreased accuracy. The supposition is that "bad" data may actually help the model generalize to new data and that removing such "noise" may not ultimately be beneficial.

4.2 Handwritten Text Segmentation

The Ancient Lives consensus locations for individual characters was the key to creating the AL-ALL and AL-PUB datasets. For this experiment I propose utilizing these locations as a way to explore the efficacy of multiple approaches for segmenting and identifying lines of handwritten text in damaged manuscripts from the Oxyrhynchus Papyri collection. Approaches similar to those demonstrated in (Barakat et al. 2021) could act as an intermediate step used to locate the text while character recognition models can then be used to classify individual characters in the segmented lines. Clustering methods, such as spectral clustering utilizing consensus locations, may also prove useful in segmenting text in these images.

4.3 Dataset Augmentation With Synthetic Instances

In (Swindall et al. 2022) it is shown that augmenting a dataset with synthetic instances can improve per-class accuracy without affecting the overall model accuracy. This leads to important questions regarding potential negative effects on un-augmented samples. The as yet understood relationship between model accuracy and per-class accuracy deserves further exploration. I propose an experiment where various samples in the AL-ALL dataset are augmented with GAN generated instances to varying degrees in an effort to understand how such a technique can be better utilized to reduce the effects of class imbalance.

Additionally I propose exploring the use of Neural Style Transfer (NST) similar to (Zhu et al. 2017) to augment samples too small for most GAN implementations. The approach here is to create handwritten instances of Sigma, the smallest sample in AL-ALL, then train the NST Discriminator on the dataset. Then the style transfer can be applied to the newly created handwritten instances to create images resembling the original data. These images can then be used to augment the original sample. Such sample augmentation may in turn be utilized by StyleGAN2 in the same manner as (Swindall et al. 2022) to further augment this sample and increase per-class accuracy.

4.4 Pipeline for Crowdsourcing Image Datasets

The final proposed experiment combines the results from each of the previous proposals into a complete pipeline for crowdsourcing image datasets. Volunteers are employed to annotate images through a web-based interface. The resulting data are then utilized to extract and label distinct samples creating a training dataset of images. The trustworthiness of the human annotations and model predictions are determined through an ensemble model approach to remove problematic and incorrectly labeled data improving model accuracy. The effects of sampling bias are reduced through augmenting small samples with synthetic instances. The new and improved dataset is then used to train classification models that can be used along side text segmentation methods for rapid transcription. To ensure that this approach can be generalized it will be necessary to apply this pipeline to a dataset other than AL-ALL. As the volunteer annotation process is similar to that used in the Ancient Lives Project, the data from the Galaxy Zoo Project may be an ideal dataset to test the proposed dataset creation pipeline.

5 Research Challenges

Many researchers incorrectly perceive that Optical Character Recognition and Text Recognition to be a solved problem, but as shown in (Swindall et al. 2021), state-of-the-art OCR engines are not capable of extracting text from damaged ancient documents or even recognizing individual handwritten characters in this context. This misconception has lead to a lack of work in this area which has resulted in a lack of useful datasets necessary to train better models. The Ancient Lives Project was a vital first step toward this goal, but uncertainty in the ground-truth in crowdsourced datasets and, in turn, uncertainty in the predictions of models trained on such datasets are complications that need to be further mitigated. Class imbalance continues to be a challenge with no clear solution and text segmentation is currently impractical for such manuscripts. The combination of these challenges has made it impossible, thus far, to develop automatic transcription systems for use in digital humanities and papyrology.

References

Barakat, B.; Droby, A.; Kassis, M.; and El-Sana, J. 2021. Text Line Segmentation for Challenging Handwritten Document Images Using Fully Convolutional Network.

Bowman, A. K.; Coles, R.; Gonis, N.; Obbink, D.; and Parsons, P. J. 2007. *Oxyrhynchus: a city and its texts*. Graeco-Roman Memoirs, v. 93. London: Published for the Arts and Humanities Research Council by the Egypt Exploration Society.

Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Russakovsky, O. 2019. Human Uncertainty Makes Classification More Robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Simpson, R.; Page, K. R.; and De Roure, D. 2014. Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*, 1049–1054. New York, New York, USA: ACM Press. ISBN 9781450327459.

Swindall, M. I.; Croisdale, G.; Hunter, C. C.; Keener, B.; Williams, A. C.; Brusuelas, J. H.; Krevans, N.; Sellew, M.; Fortson, L.; and Wallin, J. F. 2021. Exploring Learning Approaches for Ancient Greek Character Recognition with Citizen Science Data. In *2021 17th International Conference on eScience (eScience)*, 128–137. IEEE.

Swindall, M. I.; Player, T.; Keener, W. A. C., Ben and; Brusuelas, J. H.; Nicolardi, F.; D'Angelo, M.; Vergara, C.; McOsker, M.; and Wallin, J. F. 2022. Dataset Augmentation in Papyrology with Generative Models: A Study of Synthetic Ancient Greek Character Images. In *The 31st International Joint Conference on Artificial Intelligence*. IJCAI-ECAI.

Williams, A. C.; Wallin, J. F.; Yu, H.; Perale, M.; Carroll, H. D.; Lamblin, A.-F.; Fortson, L.; Obbink, D.; Lintott, C. J.; and Brusuelas, J. H. 2014. A computational pipeline for crowdsourced transcriptions of Ancient Greek papyrus fragments. In *2014 IEEE International Conference on Big Data (Big Data)*, 100–105. IEEE.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.

6 Supplement

As a researcher in a niche field, it is important to share the challenges faced with other researchers in adjacent fields, and to seek new ideas from disparate areas of study. Such work would be impossible without extensive, interdisciplinary collaboration. Many machine learning and AI researchers are unaware of the challenges faced by those who study ancient documents, and many in the digital humanities lack the technical knowledge to explore state-of-the-art solutions. HCOMP is the ideal venue to explore what is possible and what is necessary to further research in crowdsourcing image datasets and accelerating transcription of ancient manuscripts. I am currently in the process of writing my doctoral dissertation proposal and would benefit immensely from the insights of experts in related and unrelated areas of study while I contemplate the direction my future work may take. While I have an amazing team of mentors and colleagues, the challenges I face in my research necessitate that I search for new ideas, new perspectives, and new solutions. My hope is that the HCOMP Doctoral Consortium will inform and inspire new approaches and new questions to explore. I Plan to defend my dissertation in May of 2024.