Self-training and Active Learning with Pseudo-Relevance Feedback for Handwriting Detection in Historical Print

Jacob Murel[1] and David Smith[1]

Northeastern University, Boston, MA, USA
j.murel@northeastern.edu
dasmith@ccs.neu.edu

Abstract. Handwritten text recognition research largely focuses on entirely handwritten documents, yet many bibliographic researchers are interested in handwriting left by readers in historical print. Due to the sparse and inconsistent appearance of handwritten annotations, compiling sufficient datasets of handwriting in print can be difficult. We propose a method for utilizing visual similarities among text exemplars for improving handwriting detection in historical print. We investigate the effect of pseudo-labeled page images on improving object detection model performance for handwriting localization across multiple exemplars of Shakespeare's First Folio. We compare differences in self-training and active learning with pseudo-labels for positive and negative-sample images, using pseudo-relevance and relevance feedback as selection methods. We find that pseudo-labels from positive and negative-sample images improve detection task performance on individual exemplars with an average precision increase of 15%. Tests on collections of multiple exemplars are less conclusive. We discuss how variations in historical print's materiality may explain these results and outline further research to investigate this matter.

Keywords: Object detection · Handwriting · Convolutional neural network · Historical print · Digital humanities

1 Introduction

Much work in handwritten text recognition (HTR) focuses on entirely handwritten documents, yet bibliographic researchers, for whom such tools are of use, are often interested in the handwritten marks left by early readers in historical print. Despite scholars' interest in handwritten reader marks, constraints on accessibility and funding limit research to a single or small handful of texts [50,7]. Given the mass digitization of library collections and importance of handwritten marks for researchers, a tool that detects and enumerates handwriting in collections is immensely valuable.

As with many machine learning models, object detection models (ODMs) require large amounts of manually annotated data in order to produce optimal

results. Producing sufficient datasets can be computationally expensive however. Moreover, handwriting in early historical print is scarce [50,7,39], with a long-tailed distribution, which results in imbalanced datasets. Studies propose several methods for improving HTR model performance, such as active learning [9,45] and transfer learning [62,37,4]. Unfortunately, none of these techniques have been applied to ODMs for handwriting in print. Moreover, none of these approaches have considered how to utilize historical print's bibliographic characteristics to improve model performance.

Unlike other image datasets, page images from early printed books share a large visual overlap. Multiple copies of a single printed text, such as the First Folio of Shakespeare's plays (1622/3), share the same print layout and formatting. Texts in the same genre also often share bibliographic features, such as typeface size, page layout, etc. Shakespeare's First Folio, for instance, shares a similar layout with seventeenth-century folios of Ben Jonson's plays. We therefore ask: How might handwriting detection models take advantage of the mass digitization of multiple exemplars of a single printed text? Does finetuning a model to one printed text improve model generalizability to other exemplars? To address such questions, we investigate how active learning via relevance feedback and self-training via pseudo-relevance feedback can effectively take advantage of historical print's bibliographic characteristics to improve handwriting detection model performance.

We further consider the effect of negative-sample images—i.e., images of print sans handwriting—on model performance, a hitherto unexplored area in scholarship. Active learning research for image classification [47,30,55], object detection [2,63,67], and HTR [9,45] exclusively addresses sampling from positive-sample images. Given the prevalence of negative-sample images in the majority of books without annotations on every page, it is worth investigating how negative-sample images may affect (even improve) model performance. How do active learning and self-training with negative-sample images compare to positive-sample images in improving model performance? We join calls to re-evaluate the importance of negative evidence in machine learning [6] by arguing that negative-sample images are comparable to positive-sample images in improving handwriting detection for bibliographic search tasks.

By investigating the comparative impact of different image subset compositions on model performance, we also address how self-training and active learning via (pseudo-)relevance feedback and pseudo labels compare to training with manually labeled images. How do manually annotated positive-sample images compare to automatically annotated positive and negative-sample images in model training? HTR [29,5,14] and object detection [21,70,19] research propose several semi-supervised learning methods with promising results. Pseudo labels are a source of distant supervision, in which machine-generated data labels are used to improve model performance [28,44,15,46]. We contribute to ongoing discussions of unsupervised and semi-supervised learning by examining how pseudo-labeled versus manually labeled data affect model performance for multiple text exemplars.

To conduct these investigations, we compile and release open-source training data from a wide array of open-access early printed books and compile test data from eleven open-access copies of Shakespeare's First Folio (approx. 10,000 images). Our experiments suggest that self-training and active learning with positive versus negative-sample images equally impact model performance. We show that training with pseudo-labels from one exemplar markedly improves model precision on that same and other individual exemplars by an average precision of 15% and recall of 5%. We also find, however, that it negatively affects model performance on a collection of multiple exemplars, decreasing precision and recall by up to, respectively, 10% and 20% in one case. We discuss how the materiality of historical print documents may explain these results. We conclude by outlining further research for investigating the degree to which historical print's materiality affect computer vision models.

2 Related Work

A large amount of research focuses on how to reduce manual annotation efforts for handwriting detection and alleviate model performance issues resulting from imbalanced datasets. Notable methods widely discussed are self-training, active learning, and data augmentation.

2.1 Active Learning

Active learning is an algorithmic technique that selects the most informative data samples for labeling in order to improve model accuracy and efficiency. Active learning aims to reduce the amount of necessary manual data annotation by helping locate the most meaningful data samples for training as opposed to manually annotating a large, random sampling of data. Studies suggest active learning markedly improves model performance for image classification [47,30,55,59,49] and handwriting recognition in OCR pipelines [9,45]. Object detection has only recently explored active learning applications with promising results. These applications, however, often involve significant modifications to model architectures, inhibiting generalization to other detection tasks and models [13,67,66].

Published active learning experiments generally take two approaches to sampling data: uncertainty-based and diversity-based methods. The first selects samples that have the least shared information with initial labeled training data so as to maximize uncertainty measures [16,22]. The second (also called distribution-based) selects diverse samples that have minimal similarities among one another with the aim of representing the whole unlabeled data distribution [49,1,59]. Both learning approaches are supervised.

2.2 Self-training

Self-training is a self-supervised learning method that uses model predictions on unlabeled data to improve model performance. An initial model is first trained

on a small set of data—in some cases, this initial training data is entirely synthetic. The model then makes predictions on a larger set of unlabeled data. A confidence-based selection method determines which samples from the unlabeled set are correctly labeled and then adds these images to the training set for further training. This is also called pseudo-labeling [28,10]. Self-training shows promising results in image classification [53,65], object detection [46], and handwritten text recognition for OCR [15,62].

Published experiments often use fixed thresholds for selecting which pseudo-labels to add to the training set [53,28]. Flexible thresholds for multi-class predictions have also been explored with positive results [68]. One study on self-training in object detection compares a selection method independent of detector confidence against more traditional highest detector confidence selection methods with promising results [46]

2.3 Data Augmentation

Synthetic data and data augmentation are similar yet distinct methods for alleviating issues that result from insufficient training data. Synthetic data refers to the creation of artificial data, as opposed to real-world data, for training models. Several studies in object detection [17,31,57] and HTR [26,23,62] examine synthetic data applications with variable results; models trained entirely or partially on manually labeled data at times perform worse, better, or equally to models trained on entirely synthetic data.

By contrast, data augmentation augments real-world training data through image transformation techniques (rotating, cropping, etc.) that create modified duplicates of manually labeled data. This expands the size and diversity of the training data by creating artificial data samples derived from pre-existing training data. Data augmentation is widely used to improve the performance of computer vision models trained on imbalanced datasets [56,48,72]. For example, recent object detection research on imbalanced datasets of medical image demonstrates data augmentation's positive impact on model performance [41,38,24]. Although the literature on data-augmentation in HTR is less extensive, recent work suggests that select data augmentation strategies (e.g., region-specific rotation and scaling) markedly improve HTR model performance [18]

2.4 Handwriting Detection

We are aware of no studies or experiments that investigate how any of the above methods affect the detection of handwritten regions in historical print using ODMs. Only two recent studies focus exclusively on handwriting detection and localization in print with ODMs [64,34]—we refer to this task as handwriting detection. We distinguish handwriting detection from HTR, the latter being the application of OCR for transcribing handwritten characters in digitized documents. The aforementioned two studies in handwriting detection focus exclusively on comparing the performance of model architectures for handwriting localization using datasets of approximately 1000 images each. In [64], YOLOv3 and Faster

R-CNN perform comparably, with Cascade R-CNN marginally outperforming both overall. By contrast, [34] shows YOLOv4 and YOLOv5 markedly outperform Faster R-CNN and DETR.

3 Method

We investigate self-training and active learning for handwriting detection models using pseudo labels obtained through (pseudo-)relevance feedback. We train one base model and then finetune it to the First Folio's print layout using subsets of model-annotated data. After this, we test model performance on single-text and multi-text datasets, the latter comprised of multiple exemplars of the same text. We evaluate model performance for bibliographic search tasks using mean average precision and recall.

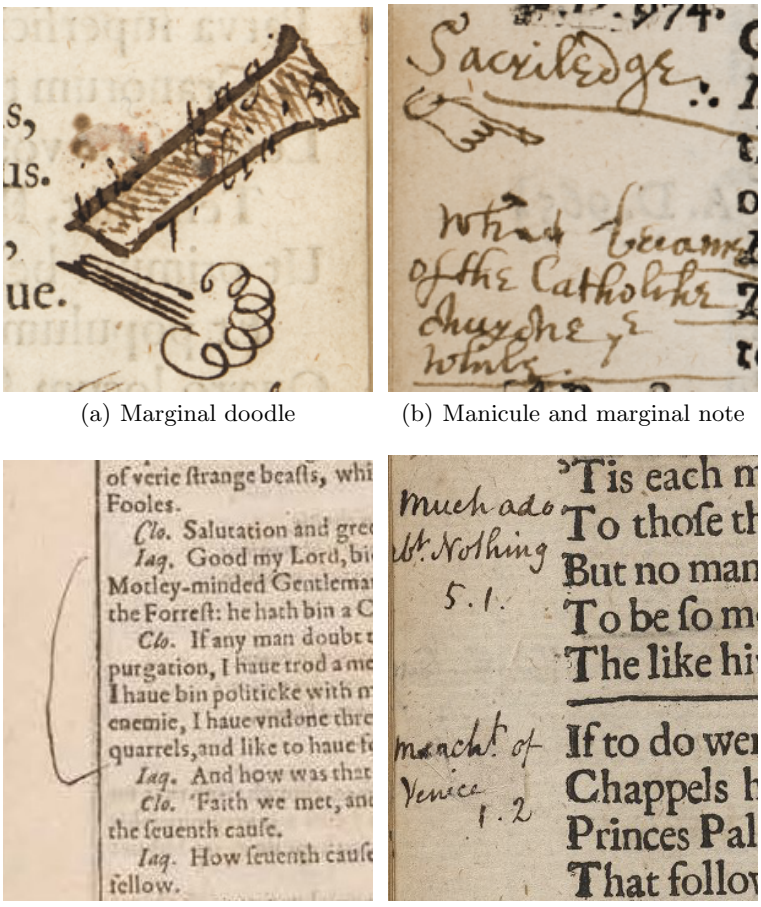
3.1 Datasets

Training Data We train our two base models using an open-access early modern reader annotations dataset.[1] The dataset is compiled from several open-access digital collections including: the Oxford University Bodleian Library, the Wellcome Library, Princeton University Library, John Carpenter Brown Library, UCLA Clark Library, the Folger Shakespeare Library's Digital LUNA Collection, Annotated Books Online, and the Munich DigitiZation Center and Bavarian State Library. While many historical print datasets focus on exclusively nineteenth-century documents [20,27,25], we compile training images from European and American books printed between the fifteenth and nineteenth centuries. The majority are in Latin type. Pages with non-Latin type—notably Arabic, Chinese, Greek, and Hebrew—constitute approximately no more than 10% percent of the total images in the dataset and are included to account for the comparably small presence of such types in early modern European and American print. Being curated from multiple institutions (which generally do not share digitization standards) page image dimensions vary. Nevertheless, all images are hi-resolution (600+ dpi) with an average image height around 1000 pixels.

The dataset contains 2,448 positive-sample images. We label all forms of ink-based handwriting, including doodles (Fig. 1(a)), manicules (Fig. 1(b)), simple brackets (Fig. 1(c)), and interlinear and marginal alpha-numerical notes (Fig. 1(d)) under one "handwriting" class. Manuscript underlining of printed text is ignored. All images are labeled by a trained paleographer using Roboflow,[2] a development tool for producing computer vision models. There are 9,830 total "handwriting" labels across all 2,448 positive-sample images. We use an equal number of negative-sample images (2,448) from this dataset for training our base models, as our past research with this dataset suggests using an equal number of positive and negative-sample images improves model accuracy [36].

[1] This dataset is publicly available for use at `https://github.com/jmurel/em_reader_ann`.

[2] `https://roboflow.com/`



(a) Marginal doodle (b) Manicule and marginal note

(c) Simple bracket (d) Alpha-numeric marginal notes

Fig. 1. Examples of handwritten marks left by early readers in historical print. (a) and (b) courtesy of UCLA Clark Library, Los Angeles, California, USA. (c) courtesy of Free Library of Philadelphia, Philadelphia, Pennsylvania, USA. (d) courtesy of Folger Shakespeare Library, Washington D.C., USA.

Test Datasets We test and train our models on digitized copies of Shakespeare's First Folio (FF). We have chosen the FF as our test text given its wide accessibility. Due to its canonical status in English literature, many FF copies have been digitized in their entirety, significantly more so than other early printed texts.[3]

[3] Sarah Werner documents forty-nine of 228 extant FFs digitized in their entirety (`https://sarahwerner.net/blog/digitized-first-folios/`)

The FF serves as a suitable test case given the scarcity of handwriting in extant copies. While most copies used in our evaluations contain some form of handwriting, only one copy contains handwriting on more than 10% of its 900+ pages. Thus, in comparison to other early printed books that may contain an abnormally high amount of handwriting (e.g., herbals or devotionals), the FF serves a suitable case study for detecting rare occurrences of handwriting in early print.

We use two data sets for model testing and active learning and self-training sample selection: a single-Folio and multi-Folio set. We use single and multi-copy sets of the same text for two principal reasons. First, we can measure how active learning with one text copy generalizes to other exemplars of the same text. Additionally, testing on one text and many texts helps account for different bibliographic search tasks. For example, in locating handwriting in printed books, researchers may want to examine only one large book or compare the proportion of pages with handwriting across several books. Additionally, library catalogers may want to determine the proportion of pages with handwriting across an entire collection or in an individual book. Such large-scale bibliographic search tasks can help illuminate how early readers interact with print—e.g. which sections of books are most frequently annotated, how prevalent handwriting may be in different copies and editions, etc. Our test sets help compare self-training methods while accounting for possible bibliographic research tasks.

The single-Folio set consists of one FF copy from the Free Library in Philadelphia, Pennsylvania, USA. We select this exemplar for two reasons. First, it is the most heavily-annotated digitized FF (containing 330 pages of handwriting) for which hi-resolution, open-access images are available. Second, because its contained handwriting has been attributed to the English poet John Milton [8,32], bibliographic scholars have meticulously documented all of its contained handwriting, providing a readily available ground truth for measuring model accuracy.

The multi-Folio test set is comprised of ten digitized FFs. We curate these from Oxford University Bodleian Library, Boston Public Library, Cambridge University Kings College Library, Dulwich College, Manchester University John Rylands Library, State Library of New South Wales, National Library of Scotland, Saint-Omer Library, Folger Shakespeare Library, the Royal Shakespeare Company, and Padova University Library.[4] We choose these ten with the intent of covering a range of manuscript annotation proportions. Some are heavily annotated while others contain nearly no handwriting.

3.2 Training

We train one YOLOv5[5] base model from our training dataset. We choose the YOLO model architecture [42] given its overall general positive performance (as

[4] Acquired through First Folios Compared: `https://firstfolios.com/view-first-folios`.

[5] `https://docs.ultralytics.com/yolov5/`

opposed to other architecture's variable performance) for handwriting detection [64,34] and HTR tasks [61,54]. Although individual image dimensions vary, no image falls under 640 pixels high, and so we adopt this size for training all of our models. We use early stopping to train the initial base model. Note that we investigate object detection rather than image classification given the small area of a printed page typically occupied by early handwriting.

We employ data augmentation for training our base model. Specifically, we utilize augmentation techniques intended to mimic early readers' bibliographic practices, as well as the effects of modern digitization processes. Specifically, we adopt augmentation techniques such as page rotation and cropping, bounding box rotation, and bounding box cut-and-pasting from one image to another.

We choose these techniques for two reasons. First, object detection [56,48,72] and HTR [18] research suggests they are effective in improving model performance. Second, these specific augmentation techniques mirror early reader practices. Early modern readers bound pages at different orientations per individual reader practices and cut-and-pasted sections from one text into another [35,52,51,60]. Later curators also often trimmed extant handwriting from page margins [7,50], and, of course, early handwriting can run in directions incongruous with printed text. We also utilize image skewing for augmentation in order to account for variations in digitization, such as camera-object positioning.

Note that we train an additional model on a modified version of our training dataset. For this additional model, we manually annotate ten page images (five positive-samples and five negative-samples) from a digitized FF, add them into our initial training data, and train a model from this modified training set. Our purpose in doing this is to create a baseline with which to compare manually annotated versus automatically annotated page images in model training.

3.3 Self-training and Active Learning

We obtain pseudo-labeled images for self-training and active learning by deploying our base model on our single-text test set. We then use relevance feedback as our active learning selection method and pseudo-relevance feedback as our self-training selection method.

Pseudo-relevance and relevance feedback are related methods for improving query results in information retrieval systems [3,33,71,12]. Using an existing retrieval system, relevance feedback selects the top n relevant results from a given query, returns them to a human user to confirm their relevance, and then uses these samples to refine the initial query. Pseudo-relevance feedback is, essentially, an automated form of relevance feedback. Whereas the latter returns the top n results from a given query to a user to confirm relevance, pseudo-relevance feedback assumes the top n results are relevant and uses them to refine the initial query. Given relevance feedback's semi-supervised process and pseudo-relevance feedback's unsupervised process, we adopt them as selection methods, respectively, for active learning and self-training.

While self-training and active learning are usually tested with out-of-sample data, (pseudo-)relevance feedback aims to improve model performance in-sample;

that is, (pseudo-)relevance feedback aims to improve retrieval accuracy for the same query and dataset provided initially by the user. Given (pseudo-)relevance feedback's use for primarily improving model performance within a dataset, we suggest it provides a useful sample selection method for measuring model generalizability across multiple exemplars of one text.

We investigate (pseudo-)relevance feedback as a selection method for positive and negative-sample images. Having deployed a base model on our single-text test set, we create six models from that base model's results. We train these six models using six separate image sets returned by the initial base model: the top ten unique positive page images, any ten unique negative page images, the top ten unique true positive page images, any ten unique true negative page images, a mix of true positive and true negative page images, and the ten highest-ranked false positives page images (with manually corrected labels).[6] With the exception of the corrected false positive set, we use the initial model's projected annotations for self-training and active learning.

Using these image sets, we create six models each from our base model. Specifically, we further train the base model with each of the given image sets for ten epochs. We then deploy each of the resulting models on both the single-text and multi-text test sets and compare their performance to that of the initial model.

By using our base model's generated pseudo labels for self-training and active learning and comparing their use to corrected false positives from the base model, as well as manually-labeled FF images, we hope to contribute to ongoing research on the comparative efficacy of supervised, semi-supervised, and unsupervised approaches for computer vision model performance.

3.4 Evaluation

We are principally concerned with how bibliographic researchers and librarians may use our handwriting detection models in searching digitized collections of historical documents for handwriting. We therefore calculate evaluation metrics appropriate for different bibliographic search tasks.

Studies often evaluate ODMs at the pixel level, using metrics such as intersection over union [64,43]. While it can be helpful to localize handwriting on the page, we propose that book-historical search tasks are better modeled as page-level retrieval tasks. We therefore evaluate model performance using mean average precision (mAP) and recall, equally common metrics for evaluating ODM performance [40].

As may be expected, our ODMs often identify multiple instances of handwriting on a single page image. We therefore take the confidence level for the highest-ranked object on a given page as the model's prediction that any handwriting exists on that page. We ignore any additional detected objects on that

[6] Throughout this paper, we use *top* to refer to those pages with the highest-ranked confidence returned by our model(s).

page. In other words, we take the detected object with the highest confidence for a given page image as representative of that page *in toto*.

We calculate mAP and recall for different tasks using our test datasets. We evaluate tasks where the researcher has selected a book and wants to locate all pages with handwriting therein. We also employ corpus-level average precision to evaluate tasks where the researcher wants to find examples of handwriting across a larger collection. These metrics, we argue, are more appropriate for search tasks where the user will not be able to examine every page of a book or every result. It is much more efficient, furthermore, to collect user feedback at the page level than by asking for individual regions to be highlighted.

Precision We calculate mAP for the top 100 unique pages returned by each model on each test set using the formula:

$$mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k \tag{1}$$

Here, AP_k is the average precision of each class k and n is the number of classes. We use 1 class, and so $n = 1$. Given we are concerned with page-level accuracy rather than object-level accuracy, we calculate mAP for the top 100 page image results returned by each ODM for the single-text and multi-text test sets.

We further calculate mAP for each individual text in the multi-text set. To do this, we count the number of true positive and total positive results for each text from the top 100 page images returned by a model and use those values to calculate individual mAPs for each text in the multi-text set. We intend this final mAP value to reflect an actual use case in which users may want to determine the likelihood or proportion of pages containing early handwriting in each book across a larger collection.

Recall Although we calculate mAP the same for each model on the single-text test sets, practical limitations demand modifications for calculating recall with our multi-text test sets. For all test sets, we use the standard recall formula:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Here, TP is the total number of true positives in the model output, and FN is the total number of false negatives.

We calculate recall for the single-Folio set using Bourne and Scott-Warren's published paleographical description of all 330 pages that contain handwriting in the Philadelphia FF [8].

To calculate recall for the multi-text set, however, we pool model outputs as is common in information retrieval evaluations. Given logistical constraints against manually combing a test set of several thousand images for reader-annotated pages, we compile a master list of every true positive image from each model's top 100 results for the multi-Folio set. We then calculate the percentage of true positive images each model identifies out of that master list.

4 Results

Model performance metrics are displayed in Tables 1 and 2, the first showing mAP for each model, the second showing recall rates for each model. *Model 1* denotes our base model used for active learning. *Pos* and *Neg* denote models self-trained with pseudo-relevance feedback using, respectively, the top-ten positive-sample images and any ten negative-sample images returned by Model 1. *TP*, *TN*, *Mix*, and *FP* denote models trained with active learning and relevance feedback using, respectively, the top-ten true-positive-sample images (*TP*), any ten true-negative-sample images (*TP*), a mix of true-positive and true-negative-sample images (*Mix*), and the top-ten ranked false-positive images with corrected labels (*FP*). *Model 2* denotes the additional base model trained using manually labeled FF page images.

Table 1. model mAP for Folio test sets

	single-Folio	multi-Folio
Model 1	.55	.92
Unsupervised		
Model 1 - Pos	.7	.82
Model 1 - Neg	.68	.86
Semi-supervised		
Model 1 - TP	.7	.8
Model 1 - TN	.69	.85
Model 1 - Mix	.67	.81
Model 1 - FP	.69	.84
Supervised		
Model 2	.59	.87

Table 1 shows model mAP values for the FF test sets. Both the base model trained without manually labeled FF images (Model 1) and the base model trained with them (Model 2) perform comparably on the single-Folio set (although Model 2 is marginally higher). Both self-training via pseudo-relevance feedback and active learning via relevance feedback with images from the single-Folio set notably improve model mAP on that set. There appears no significant difference on model precision between using all positive, negative, corrected false positive, or a mix of images from the single FF set for self-training or active learning.

Interestingly, an opposite trend in mAP scores occurs with respect to the multi-Folio set. Model 1, by far, has the highest mAP for the multi-Folio set despite having no FF images in its training data. All of the other models use annotated FF images, whether pseudo-labels or Model 2's manual labels. All of these latter models return lower mAP scores compared to Model 1.

Table 2. model recall for Folio test sets

	single-Folio	multi-Folio
Model 1	.2	.82
Unsupervised		
Model 1 - Pos	.25	.78
Model 1 - Neg	.23	.78
Semi-supervised		
Model 1 - TP	.25	.75
Model 1 - TN	.24	.76
Model 1 - Mix	.22	.73
Model 1 - FP	.24	.75
Supervised		
Model 2	.22	.69

Table 2 shows recall rates for models tested on the FF sets. The recall rates display similar trends as the mAP rates. Models produced through self-training and active learning with images from the single-Folio set score higher recall on that set but lower recall on the multi-Folio set when compared to Model 1. Moreover, once again, there is no statistically significant difference between using all positive, negative, corrected false positive, or a mix of images from the single FF set for self-training or active learning.

Self-trained models return comparable mAP and recall scores to active learning models when tested on the single-Folio set. With respect to the multi-Folio set, self-trained models demonstrate marginally less decrease in mAP and recall compared to active learning models.

5 Discussion

Both model mAP and recall scores suggest there is no significant difference on model performance between using positive and negative images for active learning and self-training. Admittedly, both tables suggest that using all (true) positives for active learning or self-training produces the greatest increase in single-Folio mAP and recall. The difference between sampling sets is around 1% absolute, however. This suggests that negative-sample and positive-sample images from the same subset may have a comparable effect on active learning and self-training for handwriting detection. We believe this points to the hitherto unexamined potential of negative-sample images for improving ODM performance. Given the prevalence of negative-sample images in imbalanced datasets, research may benefit from further examination into how negative-sample images can aid model performance.

Moreover, our tests suggest there is little difference between using correct pseudo labels (active learning) versus a mix of correct and erroneous pseudo labels (self-training) for single-text detection tasks. More interestingly, however,

self-trained models with pseudo-relevance feedback score higher mAP and recall on the multi-Folio set than active learning models. Given their unsupervised training process, the self-trained models contain a few erroneous pseudo labels, as opposed to the semi-supervised active learning models that use only correct pseudo labels. Self-training's positive impact on performance with the multi-Folio set (compared to active learning) may suggest that using select erroneous pseudo labels positively impacts model generalizability for handwriting detection.

The most surprising trend in our experiments is the inverse trend in model performance between the single and multi-Folio test sets. That is, self-training and active learning with (pseudo-)relevance feedback universally improves model performance on the single-Folio set while impairing performance on the multi-Folio set. Does this mean finetuning a model to one exemplar of a text only increases a model's performance on that same exemplar and not other copies? To illuminate this matter, we conduct additional tests.

Table 3. model mAP for additional test sets

	Philadelphia FF	Padua FF	Omer FF	multi-Folio
Model 1	.55	.77	.39	.92
Unsupervised				
Model 1 - Phil Pos	.7	.81	.52	.83
Model 1 - Padua Pos	.75	.85	.52	.86
Model 1 - Omer Pos	.75	.84	.49	.85
Model 1 - Multi Pos	.72	.85	.58	.90
Semi-supervised				
Model 1 - Phil TP	.7	.81	.54	.8
Model 1 - Padua TP	.75	.85	.52	.86
Model 1 - Omer TP	.74	.82	.55	.86
Model 1 - Multi TP	.74	.84	.55	.87

We perform self-training and active learning via (pseudo-)relevance feedback using two separate FF exemplars from the multi-Folio set: the University of Padua FF and Saint Omer Library FF. Having deployed our base model (Model 1) on these two exemplars, we gather pseudo labels for the top positive and true-positive results from each text to produce four additional models: two with self-training, two with active learning. We also take the top ten positive and true-positive page images returned by our base model from the multi-Folio set and train two new models via self-training and active learning with those images. We then test each of these models on the three individually sampled FFs (Padua, Omer, and Philadelphia) and the multi-Folio set. Table 3 compares these new models' mAP scores to those of Model 1 and the previous models finetuned to the Philadelphia FF. Regarding the latter, we here rename *Model 1 - Pos* and *Model 1 - TP* as *Model 1 - Phil Pos* and *Model 1 - Phil TP* respectively to highlight that those models are trained using images from the Philadelphia FF.

All models finetuned to an individual FF via self-training or active learning show significant increase in model mAP for each individual FF—approximately 15-20% for the Philadelphia FF, 5-8% for the Padua FF, and 10-19% for the Omer FF. Our additional tests suggest that active learning and self-training with automatically labeled images from any FF improve model precision on other individual FFs. We believe this evidences that finetuning to one exemplar can improve model performance on other individual exemplars of the same text.

Indeed, this is bolstered by scores returned from *Model 1 - Multi Pos*. This model, produced by self-training with positive-sample images via pseudo-relevance feedback from the multi-Folio set, reveals significant increase in precision for each individual FF and only .02 loss of precision for the multi-Folio set. Self-training with pseudo labels from multiple exemplars may thus improve model performance on individual examples while retaining model generalizability.

Nevertheless, our additional tests reaffirm self-training and active learning's overall ostensibly negative impact on model precision per the multi-Folio set. Why does self-training and active learning with individual FF page images negatively affect model performance on the multi-Folio set? We suggest this may be explained by regression to the mean. That is, Model 1 already performs with such high precision on the multi-Folio set that additional finetuning negatively impacts its performance on that set. This may also point to a higher learning curve necessary for multi-text search tasks. Further experiments are needed to confirm this however.

Interestingly, in Table 3, the models produced using pseudo-labels from the Philadelphia FF perform the worst on that same FF while other models trained using specific FF sets all score highest for their respective FFs—e.g., the model produced through active learning with the Padua FF performs the best on that FF, and so forth (the exception being *Model 1 - Multi Pos*'s score on the Omer set). We suggest this may be attributed to the nature of reader marks in the Philadelphia FF. Only 10% of handwritten marks in that FF are alphabetic, the rest are simple brackets (Fig. 1(c)). Moreover, all of positive-sample images used for self-training and active learning from this FF are alphabetic, and none of the other FFs contain similar handwritten simple brackets. Thus, none of the image subsets used for self-training or active learning, even when taken from the Philadelphia FF, reflect the nature of most handwritten marks found in the Philadelphia FF. We propose this evidences how the paleographical characteristics of reader marks (in addition to print's bibliographic features) can affect model performance.

6 Conclusion

We investigate self-training and active learning with (pseudo-)relevance feedback for handwriting detection on historical print. Specifically, we experiment with different compositions of fine-tuning data. These experiments show that training with pseudo labels from one text exemplar improves model performance on that same and other exemplars by approximately 15% precision and 5% recall. These

experiments further suggest that it negatively affects model performance on a single collection of multiple exemplars, decreasing precision and recall, respectively, by up to 10% and 20%.

7 Future Work

Handwriting detection with historical print must account for both the bibliographic features of print documents and paleographic features of handwriting therein. We believe our experiments demonstrate the need for further research on how historical print's materiality and differences in digitization pipelines affect computer vision models, whether for handwriting detection or HTR.

Multiclass labeling may serve as one potential avenue for such research. Studies widely testify to how image and label classes are interrelated and correlated [58,69,11]. The presence of one object can, at times, be used to predict the presence and location of another object. Labeling print and handwritten regions in our dataset may improve model performance with respect to the latter, particularly considering many false positives in our tests are print features (e.g., italicized speaker labels, page numbers, etc.). Moreover, comparing ODM performance regarding print versus handwriting label classes may help measure the comparative impact of bibliographic versus paleographic features on model performance.

References

1. Agarwal, S., Arora, H., Anand, S., Arora, C.: Contextual diversity for active learning. In: European Conference on Computer Vision. pp. 137–153 (2020)
2. Aghdam, H.H., Gonzalez-Garcia, A., Weijer, J.v.d., Lopez, A.M.: Active learning for deep detection neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3672–3680 (2019)
3. Ahmed, A.: Implementing relevance feedback for content-based medical image retrieval. IEEE Access **8**, 79969–79976 (2020)
4. Aradillas Jaramillo, J.C., Murillo-Fuentes, J.J., M. Olmos, P.: Boosting handwriting text recognition in small databases with transfer learning. In: 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 429–434 (2018)
5. Ball, G.R., Srihari, S.N.: Semi-supervised learning for handwriting recognition. In: 10th International Conference on Document Analysis and Recognition. pp. 26–30 (2009)
6. Borji, A.: Negative results in computer vision: A perspective. Image and Vision Computing **69**, 1–8 (2018)
7. Bourne, C.: Marking shakespeare. Shakespeare **13**, 367–386 (2017)
8. Bourne, C., Scott-Warren, J.: "thy unvalued Booke": John Milton's Copy of the Shakespeare First Folio. Milton Quarterly **56**, 1–85 (2002)
9. Burdett, E., Fujimoto, S., Brown, T., Shurtz, A., Segrera, D., Sorenson, L., Clement, M., Price, J.: Active transfer learning for handwriting recognition. In: Proceedings of the 18th International Conference on Frontiers in Handwriting Recognition. p. 245–258 (2022)

10. Cascante-Bonilla, P., Tan, F., Qi, Y., Ordonez, V.: Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI). vol. 35, pp. 6912–6920 (2021)
11. Cevikalp, H., Benligiray, B., Nezih Gerek, O., Saribas, H.: Semi-supervised robust deep neural networks for multi-label classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 9–17 (2019)
12. Chen, J., Ma, R., Su, Z.: Weighting visual features with pseudo relevance feedback for cbir. In: Proceedings of the ACM International Conference on Image and Video Retrieval. p. 220–227 (2010)
13. Choi, J., Elezi, I., Lee, H.J., Farabet, C., Alvarez, J.M.: Active learning for deep object detection via probabilistic modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10264–10273 (October 2021)
14. Fogel, S., Averbuch-Elor, H., Cohen, S., Mazor, S., Litman, R.: Scrabblegan: Semi-supervised varying length handwritten text generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4324–4333 (2020)
15. Frinken, V., Bunke, H.: Self-training for handwritten text line recognition. In: Proceedings of the 15th Iberoamerican Congress Conference on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. p. 104–112 (2010)
16. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning. p. 1183–1192 (2017)
17. Gastelum, Z.N., Shead, T., Higgins, M.: Synthetic training images for real-world object detection (2020)
18. Heil, R., Breznik, E.: A study of augmentation methods for handwritten stenography recognition. In: Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA). pp. 134–145 (2023)
19. Hua, W., Liang, D., Li, J., Liu, X., Zou, Z., Ye, X., Bai, X.: Sood: Towards semi-supervised oriented object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15558–15567 (2023)
20. Im, C., Kim, Y., Mandl, T.: Deep learning for historical books: Classification of printing technology for digitized images. Multimedia Tools and Applications **81**(4), 5867–5888 (2022)
21. Jeong, J., Lee, S., Kim, J., Kwak, N.: Consistency-based semi-supervised learning for object detection. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32 (2019)
22. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2372–2379 (2009)
23. Kang, L., Rusinol, M., Fornes, A., Riba, P., Villegas, M.: Unsupervised writer adaptation for synthetic-to-real handwritten word recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3502–3511 (2020)
24. Kim, C., Kim, G., Yang, S., Kim, H., Lee, S., Cho, H.: Chest x-ray feature pyramid sum model with diseased area data augmentation method. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2757–2766 (October 2023)

25. Kim, Y., Mandl, T., Im, C., Schmideler, S., Helm, W.: Applying computer vision systems to historical book illustrations: Challenges and first results. In: Proceedings of the 5th Conference Digital Humanities in the Nordic Countries. pp. 255–260 (2020)
26. Krishnan, P., Jawahar, C.V.: Hwnet v2: an efficient word image representation for handwritten document. International Journal on Document Analysis and Recognition **22**, 387–405 (2019)
27. Kusetogullari, H., Yavariabdi, A., Hall, J., Lavesson, N.: DIGITNET: A Deep Handwritten Digit Detection and Recognition Methods Using a New Historical Handwritten Digit Dataset. Big Data Research (23), 1–13 (2021)
28. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML Workshop on challenges in representation learning (2013)
29. Leifert, G., Labahn, R., Sánchez, J.A.: Two semi-supervised training approaches for automated text recognition. In: 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 145–150 (2020)
30. Liu, Z.Y., Huang, S.J.: Active sampling for open-set classification without initial annotation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 4416–4423 (July 2019)
31. Ljungqvist, M.G., Nordander, O., Skans, M., Mildner, A., Liu, T., Nugues, P.: Object detector differences when using synthetic and real training data **4** (2023)
32. McDowell, N.: Reading Milton reading Shakespeare politically: what the identification of Milton's First Folio does and does not tell us. The Seventeenth Century **36**(4), 509–525 (2021)
33. Meilhac, C., Nastar, C.: Relevance feedback and category search in image databases. In: Proceedings IEEE International Conference on Multimedia Computing and Systems. vol. 1, pp. 512–517 (1999)
34. Moustapha, M., Tasyurek, M., Ozturk, C.: A Novel YOLOv5 Deep Learning Model for Handwriting Detection and Recognition. International Journal on Artificial Intelligence Tools **32**(4), 1–33 (2023)
35. Murel, J.: On the full- and half-sheet inserts of andreas vesalius's de humani corporis fabrica (1543 and 1555). The Papers of the Bibliographical Society of America **115**(1), 43–65 (2021)
36. Murel, J., Smith, D.: Detecting manuscript annotations in historical print: Negative evidence and evaluation metrics. In: Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM). vol. 13, pp. 745–752 (2024)
37. Nair, R.R., Sankaran, N., Kota, B.U., Tulyakov, S., Setlur, S., Govindaraju, V.: Knowledge transfer using neural network based approach for handwritten text recognition. In: 13th IAPR International Workshop on Document Analysis Systems (DAS). pp. 441–446 (2018)
38. Nguyen-Mau, T.H., Huynh, T.L., Le, T.D., Nguyen, H.D., Tran, M.T.: Advanced augmentation and ensemble approaches for classifying long-tailed multi-label chest x-rays. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2729–2738 (October 2023)
39. Orgel, S.: Margins of truth. In: Murphy, A. (ed.) The Renaissance text: Theory, editing, textuality, pp. 91–107. Manchester University Press (2013)
40. Padilla, R., Netto, S.L., da Silva, E.A.B.: A survey on performance metrics for object-detection algorithms. In: International Conference on Systems, Signals and Image Processing. pp. 237–242 (2020)

41. Rao, A., Lee, J.Y., Aalami, O.: Studying the impact of augmentations on medical confidence calibration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2462–2472 (October 2023)
42. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016)
43. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
44. Rizve, M.N., Duarte, K., Rawat, Y.S., Shah, M.: In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In: International Conference on Learning Representations (ICLR) (2021)
45. Romero, V., Sánchez, J.A., Toselli, A.H.: Active learning in handwritten text recognition using the derivational entropy. In: Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition. pp. 291–296 (2018)
46. Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-supervised self-training of object detection models. In: Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05). vol. 1, pp. 29–36 (2005)
47. Saifullah, S., Agne, S., Dengel, A., Ahmed, S.: Analyzing the potential of active learning for document image classification. International Journal on Document Analysis and Recognition **26**, 187–209 (2023)
48. Saini, M., Susan, S.: Tackling class imbalance in computer vision: A contemporary review. Artificial Intelligence Review **56**, 1279–1335 (2023)
49. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations (2018)
50. Sherman, W.: What Did Renaissance Readers Write in Their Books? In: Andersen, J., Sauer, E. (eds.) Books and Readers in Early Modern England: Material Studies, pp. 119–137. University of Pennsylvania Press, Philadelphia (2002)
51. Sherman, W.: Used Books: Marking Readers in Renaissance England. University of Pennsylvania Press (2009)
52. Smyth, A.: Material Texts in Early Modern England. Cambridge University Press (2018)
53. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: Advances in Neural Information Processing Systems. vol. 33, pp. 596–608. Curran Associates, Inc. (2020)
54. Tan, Y.F., Connie, T., Goh, M.K.O., Teoh, A.B.J.: A pipeline approach to context-aware handwritten text recognition. Applied Sciences **12**(4) (2022)
55. Tang, Y.P., Huang, S.J.: Self-paced active learning: Query the right thing at the right time. In: Proceedings of the AAAI Conference on Artificial Intelligence. p. 5117–5124 (July 2019)
56. Thabtah, F., Hammoud, S., Kamalov, F., Gonsalves, A.: Data imbalance in classification: Experimental evaluation. Information Sciences **513**, 429–441 (2020)
57. Vanherle, B., Moonen, S., Van Reeth, F., Michiels, N.: Analysis of training object detection models with synthetic data. In: 33rd British Machine Vision Conference BMVC (2022), `https://bmvc2022.mpi-inf.mpg.de/0833.pdf`
58. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: Cnn-rnn: A unified framework for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2285–2294 (2016)

59. Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L.: Cost-effective active learning for deep image classification. IEEE Transactions on Circuits and Systems for Video Technology **27**(12), 2591–2600 (2017)
60. Watt, T.: Cheap Print and Popular Piety, 1550–1640. Cambridge University Press (1993)
61. Wen, Y., Ke, W., Sheng, H.: Improved handwritten numeral recognition on mnist dataset with yolo and lstm. In: 6th International Conference on Universal Village. pp. 1–5 (2022)
62. Wolf, F., Fink, G.: Combining self-training and minimal annotations for handwritten word recognition. In: Proceedings of the 18th International Conference on Frontiers in Handwriting Recognition. p. 300–315 (2022)
63. Wu, J., Chen, J., Huang, D.: Entropy-based active learning for object detection with progressive diversity constraint. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9397–9406 (2022)
64. Wu, Y., Hu, Y., Miao, S.: Object Detection Based Handwriting Localization. In: ICDAR 2021 Workshop: Industrial Applications of Document Analysis and Recognition. pp. 225–239 (2021)
65. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
66. Yang, C., Huang, L., Crowley, E.J.: Plug and play active learning for object detection. arXiv preprint arXiv:2211.11612 (2022)
67. Yuan, T., Wan, F., Fu, M., Liu, J., Xu, S., Ji, X., Ye, Q.: Multiple instance active learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5330–5339 (June 2021)
68. Zhang, B., Wang, Y., Hou, W., WU, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In: Advances in Neural Information Processing Systems. vol. 34, pp. 18408–18419 (2021)
69. Zhang, J., Wu, Q., Shen, C., Zhang, J., Lu, J.: Multilabel image classification with regional latent semantic dependencies. IEEE Transactions on Multimedia **20**(10), 2801–2813 (2018)
70. Zhou, Q., Yu, C., Wang, Z., Qian, Q., Li, H.: Instant-teaching: An end-to-end semi-supervised object detection framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4081–4090 (2021)
71. Zhou, Z.H., Chen, K.J., Dai, H.B.: Enhancing relevance feedback in image retrieval using unlabeled data. ACM Transactions on Information Systems **24**(2), 219–244 (2006)
72. Zoph, B., Cubuk, E., Ghiasi, G., Lin, T.Y., Shlens, J., Le, Q.: Learning data augmentation strategies for object detection. In: uropean Conference on Computer Vision. p. 566–583 (2020)