Handwriting Recognition with Novelty: Supplemental Material*

Derek S. Prijatelj[1][0000−0002−0529−9190], Samuel Grieggs[1][0000−0002−2433−5257], Futoshi Yumoto[2][0000−0001−5354−8254], Eric Robertson[2][0000−0002−4942−4619], and Walter J. Scheirer[1][0000−0001−9649−8074]

[1] University of Notre Dame, Notre Dame IN 46556, USA
{dprijate, sgrieggs, walter.scheirer}@nd.edu
[2] PAR Government, 421 Ridge St, Rome NY 13440, USA
{futoshi_yumoto, eric_robertson}@partech.com

1 Supplemental Material

This is the supplemental material for the main paper intended to provide complete details of the Handwriting Recognition (HWR) domain and agent-centric approach to it for others interested in working on this challenge problem. Additional notes are provided for the domain formalization and the evaluation protocol. A large selection of supplemental experiments is provided to explore different aspects of the HWR domain with novelty in more detail.

2 Formalization of HWR with Novelty

This section consists of further details on the formalization of the HWR domain. Specifically, we provide more discussion on the HWR novelty theory, oracle definition, and ontology here.

2.1 A Theory Novelty for HWR: Additional Details

The theory of novelty for the HWR domain in the main paper notably excludes two pieces from Boult et. al [1]. One of those components is the agent's (α) action space $\mathcal{A}$ that contains all possible actions $a_t \in \mathcal{A}$ that the agent may take. The other part is the state recognition function $f_t(x_t, s_t) : \mathbb{R}^d \times \mathcal{S} \mapsto \mathcal{S} \times \mathcal{A}$, where x_t is an observation-space input, and $s_t \in \mathcal{S}$ is the agent's state at the current time-step t for all possible states $\mathcal{S}$ of the agent. In traditional image classification

* This research was sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under multiple contracts/agreements including HR001120C0055, W911NF-20-2-0005,W911NF-20-2-0004,HQ0034-19-D-0001, W911NF2020009. The views contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the DARPA or ARO, or the U.S. Government.

tasks, these two components do not come into play as the agent is neither stateful nor does it take actions in the world beyond outputting the predicted class for a sample. In image classification, as well as the HWR domain with novelty as defined in the main paper, the state recognition function could simply be the predicted class for a given sample.

However, these two components could be a part of an agent for the HWR domain with novelty. For example, the state of the agent could update if the agent were to include incremental learning to solve one of the HWR tasks. The incremental learning would then be a state change in the learning for the agent, as would any "learning" that were to occur over time, such as the continued training of an artificial neural network. An extreme example of incremental learning would be an agent that is pre-trained on available handwritten documents, but is given a new set of documents in a different language, thus with new glyphs, characters, words, language, etc., to be learned and is only given these new documents over time as they are discovered.

If the open world HWR agent proposed in the main text were to be extended and given the ability to act in the world, then the possible actions $\mathcal{A}$ would come into play. An example of this would be an automated robot performing the entire handwritten document transcription process itself, *e.g.*, opening the physical books carefully, gently turning the pages, and possibly changing the perceptual operator by zooming in and out of the images or changing sensors.

2.1.1 Caveats of Defining the Oracle The oracle mainly consists of the datasets used for evaluation. However, the oracle also determines world dissimilarity and regret functions given the data. For datasets that consist of ground truth labels typically used in supervised learning, the datasets may be enough. However, there may be additional domain knowledge that is not available in the datasets, either implicitly or explicitly defined by the oracle. An example is the case of label frequency within the dataset mismatching the current domain knowledge of the world. In this case, the oracle provides label weightings to adjust the dataset's samples to better represent the current domain knowledge of the universal population of those labels. The information about these weightings may or may not be provided to the agent as an extended part of the world space.

Another example case is when the dataset has missing labels, either partially or completely, the oracle is expected to provide the information that defines the task given the data. A complete lack of labels for a certain type of information is a rather common occurrence in domains where novelty is present and cannot be labeled in any capacity, and where labeled or unlabeled data may be used in training and evaluation. The datasets do not necessarily define the oracle's information in its entirety. This is a challenge for evaluation design that needs to be accounted for when assessing agents that must manage novelty. A sampling problem, such as HWR, thus defines the oracle and task through the world space, world dissimilarity, world regret, and the information used from datasets and domain knowledge.

2.2 HWR Ontological Specification

An ontological specification serves to characterize the novelty space and provide a basis for measuring the difficulty of detecting novelty within that space. Novelty in HWR is organized by three categories: writing style, pen selection and background novelties, which are the novelties that are not specific to the content of the text. The intent of specification is to describe all observable novelties (from an oracle's view of the world) including environmental novelties such as water damage to the writing medium, temporal and locale novelties such as date and time representations and document structures, and text-related novelties such as copyedit marks.

In the development of the ontology-based knowledge graph, we first start with the characterization of writing style. Writing style is made up of the style attributes slant angle, word space, character size and pen pressure. Each style attribute is described by a continuous function, defined in Table 1, which is applied to images of words present in each writing sample [3]. The results from the functions for all samples are binned to form discrete style descriptors, which are used to construct style attribute nodes in the knowledge graph. The number of bins is chosen to provide adequate separation of each writing style. In our initial assessment, we used four bins for slant angle and three bins for the rest of the style attributes. The style attributes are collected for all writing sample images. The most frequent style attribute value is assigned to each writer. The result is a set of associations between each writing sample, the style, and the writer, as shown in the knowledge graph in Fig. 1. We apply the same approach for background and pen novelties. The non-style measurement functions for these novelties are described in Table 2.

Style	**Function**
Pen Pressure	$\left(\sum_{i=1}^{\mathcal{N}} pixel[i]\right) / \mathcal{N}$ where $\mathcal{N}$ is the number pixels in the written text, and $pixel[i]$ is the intensity of a pixel i.
Slant Angle	$\max_{A^i} S(A^i)$ where A^i is the set of angles [-45,-30,-20,-15,-5,0,5,15,20,30,45], and $S(A^{(i)})$ is a shear estimate [9].
Word Spacing	Average number of horizontal pixels between words where a space is a vertical slice with fewer than 30% quantile of vertical pixels for a line image.
Character Size	Average number of pixels over all vertical slices of the image excluding those slices labeled as a space.

Table 1: Style Measurement Functions

Novelty Type	Function
Background	Entropy of the grey level background (without text)
Pen	Entropy of the grey level pixel intensities in the written text.

Table 2: Non-Style Measurement Functions.

A correct knowledge graph consists of each writing sample associated to a single writer via a two step path through the four style attribute nodes. The writing style measurement functions provide a gross measure of writing style. Combined with the inaccuracies introduced with binning, not all writing samples from the same writer are associated with the same set of bins across all style attributes. Since the same writer's style is an aggregate value over a set of writing samples, some binned measures for a writing sample form an association to a style attribute not associated with the writer of the sample. This is highlighted in the sample graph in Fig. 1 via a red edge. This suggests an optimization strategy for style binning and association to maximize the number of writing samples associated with a writer through the four style attribute nodes.

2.3 Ontological Specification for Novelty Characterization

Characterization was achieved through groups of clusters over writer samples created by the agent. Each group explains a single characterization of novelty as it occurs in each text image. Groups included in our initial study are:

- Up to 3 clusters for pen pressure, character size and word spacing,
- Up to 4 clusters for slant angle,
- Up to 3 clusters for category of novelty: writer novelty, background and pen novelties.

A single 'writer novelty' cluster occurs in the novelty category cluster group when novelty does not occur — all non-novel examples cluster together. Fig. 2 illustrates this approach with two cluster groups.

For performance evaluation of characterization, we use Normalized Mutual Information (NMI) to measure the quality of the clusters. We first separate the agent characterizations of writing samples with no novelty and the three categories of novelty: writing style, pen and background. We interpret characterizations in the non-novel subgroup as a base measurement of the agent's dependence on the cluster-represented attributes to describe novelty.

The characterization promotes better understanding of an agent's performance in the HWR domain with novelty. We first establish a baseline cluster quality using non-novel writing samples. In the baseline, cluster groups organize samples by similar styles and backgrounds. As different types of novelty are introduced, new cluster centers are formed to isolate those samples perceived as having the group's representative novelty.

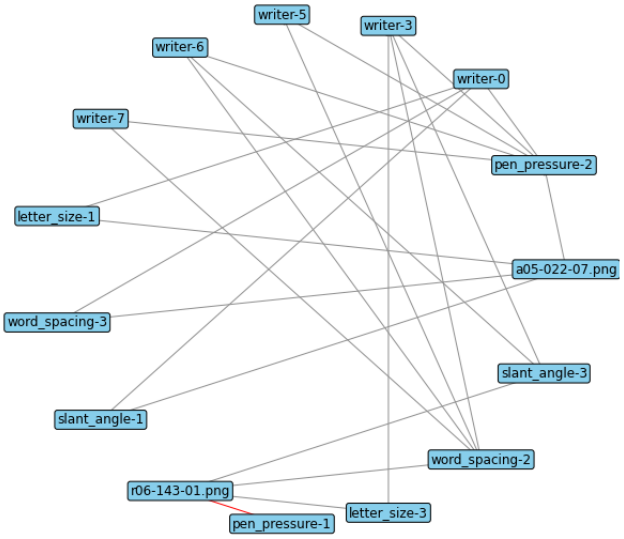

Fig. 1: Illustrative Knowledge graph of Writing Style for four style attributes associated writing samples a05-022-07 and r06-143-01 and five selected writers. The red edge on the bottom represents the writing style of a sample not associated with the sample's author.

We partition and evaluate the characterization clusters by novelty category, shown in each row of Table 3, to highlight the interactions between different categories of novelty and the novel style attributes. Applicable measures to this structure include NMI and cluster purity. This structure for analysis aids in understanding agent response to mixed novelties, such as style and background changes. The No Novelty row serves as a baseline characterization of writing samples without novelty. The Style row measures characterization clusters of samples with novel writing styles. In terms of a mapping to empirical observations, low performance for cell PP_p (Pen, Pen Pressure) in comparison to

Novelty	**PP**	**CS**	**WS**	**SA**	**NC**
Style	PP_s	CS_s	WS_s	SA_s	NC_s
Background	PP_b	CS_b	WS_b	SA_b	NC_b
Pen	PP_p	CS_p	WS_p	SA_p	NC_p
No Novelty	PP_n	CS_n	WS_n	SA_n	NC_n

Table 3: Characterization cluster groups are Pen Pressure (PP), Character Size (CS), Word Spacing (WS), Slant Angle (SA), and Novelty Category (NC).

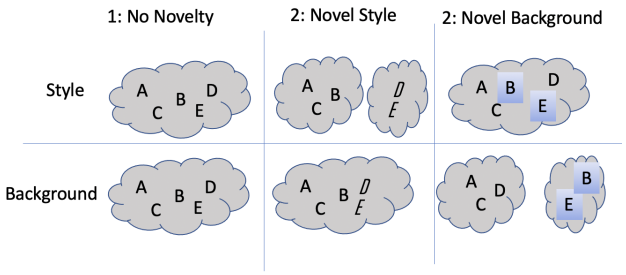

Fig. 2: Example Clusters for two cluster groups, Style and Background, under three novelty types: (1) No Novelty; (2) Novel Style; and (3) Novel Background.

baseline No Novelty was observed, indicating that this paper's open world agent is unable to discern pen changes from pen pressure novelty. We do not expect to see much variation from the non-novel examples in cells CS_p, WS_p and SA_p, since pen pressure novelties do not significantly affect character size, slant angle and word spacing. We also expect matched performance to the baseline No Novelty conditions for separable novelty categories, such as all Style cells in the Background row (PP_b, CS_b, WS_b, SA_b). For example, an agent is expected to separate novel from non-novel backgrounds, but may fail to adequately separate groups of samples using two different novel backgrounds.

The Novelty Category (NC) cluster group serves to characterize the core types of novelty. NC cells NC_c, NC_b and NC_p are meant to be measurements of an agent's ability to distinguish different samples within the same category of novelty. For example, NC_b measures an agent's ability to distinguish examples with blue backgrounds from those with red backgrounds.

For an initial assessment, we characterized the last 32 test images selected from each test prior to evaluating characterization of the novelty. We provide the sample set of measurements using Cluster Purity in Table 4.

$$\text{Purity} = \frac{1}{N} \sum_{i=1}^{k} max_j |c_i \cap t_j| \tag{1}$$

where N = number of writing samples, k = number of clusters, c_i is a cluster in C, and t_j is a ground truth novelty label.

In this sample, we see evidence of confounding variables when characterizing pen pressure with pen changes. Characterization of slant angle, when compared to non-novel cases, is weakly affected by pen changes. Word spacing was significantly affected in all three novelty cases. Style changes were correctly separated from background and pen novelties, as indicated in the Novelty Category cluster group.

Novelty	PP	LS	WS	SA	NC
Style	0.88	0.85	0.55	0.53	1.00
Background	0.83	0.89	0.45	0.61	0.89
Pen	0.75	0.71	0.57	0.77	1.00
Non-Novel	0.84	0.75	0.80	0.80	1.00

Table 4: Cluster Purity Characterization Results based on novelty type. Characterization cluster groups here are Pen Pressure (PP), Letter Size (LS), Word Spacing (WS), Slant Angle (SA) and Novelty Category (NC).

3 Additional Information on HWR Agents

The details of the baseline open world HWR agent generally defined in the main paper are included here, along with an additional agent that is not designed to handle novelty.

3.1 Baseline Open World HWR Agent

This paper's proposed HWR novelty detecting agent is further specified in this section. This baseline open world agent for HWR is built upon the Convolutional Recurrent Neural Network (CRNN) architecture which is commonly used for closed set HWR tasks. The IAM dataset [4], a very commonly used handwritten text dataset, contains a number of writing errors that were introduced when the dataset was created. These are mostly in the form of crossed out misspelled words. The ground truth provided with the dataset represents these errors, with the "#" character. This is treated as the baseline agent's exposure to novel characters, serving as the known unknown class. This known unknown class was further expanded upon by introducing a subset of novel characters from the RIMES dataset, which contains numerous characters with diacritics that are not present in IAM.

For all experiments using a CRNN, the model was structured as follows and is shown in Table 6. Five convolutional layers feed into five bidirectional LSTM layers, each with a kernel size of 3 and a stride and padding of 1. The 5 LSTM layers have a hidden size of 256. Input Images were resized to 64 pixels tall. The CRNNs were trained until they did not improve for 80 epochs using the RMSprop optimizer with an initial learning rate of $3 * 10^{-4}$. On average the models would train for around 300 epochs, using a batch size of 8. Training proceeded indiscriminately on a selection of Nvidia Titan X, Titan Xp, 1080ti, 2080ti and RTX 6000 GPUs, with each epoch taking about 5-10 minutes depending on the GPU used. Inference averaged about 33 milliseconds per sample on a 2080ti.

The CRNN serves as both the feature extractor and transcript predictor as seen in the main paper's Fig. 2. In the supplemental Figure 3, the CRNN portion of the agent is depicted in isolation to indicate its merging of feature extraction and transcript predicting during its training in a supervised learning fashion. To be used as a feature extraction for the EVMs for writer identification and

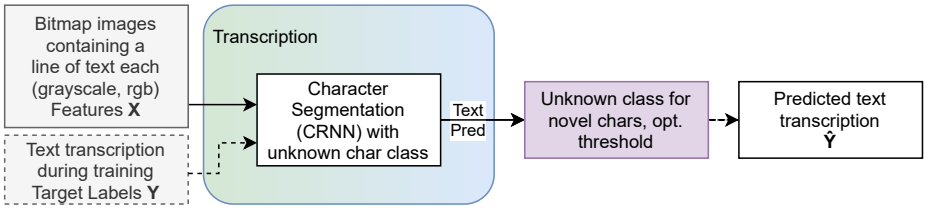

Fig. 3: The CRNN in isolation depicting its joint use as both the feature extractor and transcript predictor. The feature extraction occurs in joint with its supervised learning of the transcription task. The penultimate layer of the CRNN is used as one of the examined feature spaces for training the style task EVMs, after using PCA on the zero padded sequential output. Due to memory constraints, 1,000 components were used with from an incremental implementation of PCA on 25% of the training data.

ODAI, the penultimate layer of the CRNN (specifically the last layer of its RNN portion), was used as the encoding of the line images. Given that the sequence differs with the length of the input line image, the encodings were padded with zeros to the maximum time-step size (656 with the IAM and RIMES data) and then an incremental principal component analysis [6] method was used with 1000 components to obtain a memory-managable encoding for the EVMs. And to expedite the process given both memory and time constraints we used only one fourth of the data after running through the CRNN. Due to using 1000 components to fit in memory, a lot of useful information was lost and probably significantly affected the performance of the MEVMs that used the CRNN-PCA as their feature space.

The Extreme Value Machine (EVM) is an open set classifier designed to handle novel classes [7]. The EVM has various hyperparameters, including tail size, cover threshold, distance measure, and distance multiplier. The tail size used was 1000, the cover threshold was 0.5, the distance measure used was Cosine similarity, and the distance multiplier was 0.5. These hyperparameters were the same for the two separate EVMs trained on their respective style tasks of writer identification and ODAI. The implementation of the EVM used is a pytorch version with GPU support[3]. EVM training took approximately 2 hours per training fold on both data sets, thus equating to 20 hours total of EVM training given the two experiments with 5-fold cross validation. Prediction time was approximately one half hour for each evaluation fold using the EVM. Hardware used for training and inference of the EVM matched that of the CRNN described above.

The EVM's output a probability vector of size $K+1$ for K known labels. The extra label serves as the general novel class label (referred to as the unknown class in the EVM documentation). To obtain this probability vector, the EVM outputs all of the probabilities for the K classes in its implementation. To obtain the probability of the novel class, the probability of the maximum probable

[3] This will be made publicly available after the publication of this paper.

zero_padding2d	(115, 115, 1)
Conv-2D	(58, 58, 32)
MaxPooling2D	(29, 29, 32)
Conv-2D	(29, 29, 64)
MaxPooling2D	(14, 14, 64)
Conv-2D	(14, 14, 128)
MaxPooling2D	(7, 7, 128)
Flatten	6,272
DropOut	6,272
Dense	512
DropOut	512
Dense	256
DropOut	256
Dense	50

Table 5: Baseline Closed World Writer-Predictor Agent Model.

known class k_m is taken with $1 - k_m$ calculated as the probability of the novel class. The rest of the known probabilities are scaled by k_m and the probability of novelty is appended to the end of the probability vector.

3.2 Baseline Closed World HWR Agents

Two additional closed world agents were evaluated as comparison points to the open world agent. One agent performs the writer identification style subtask, while the other performs the text transcription task. They do not pass information between each other, and have no specific abilities to manage novelty.

Baseline Closed World Writer Identification Agent A baseline closed world agent for just writer identification was created for comparison to the open world agent, described in the main text, under novel conditions. The closed world baseline agent predicts, for each sample, one of the 50 known writers in the training set by applying the softmax function to the output of the dense layer of a CNN. The baseline model serves to demonstrate limited utility only in a closed world, over-fit to known writers, with considerable degradation in performance when exposed to novel conditions.

The baseline writer identification model is a neural network consisting of three groups of 2-D convolution layers with RELU activation and max pooling, followed by two groups dense connected layers with RELU activation and 50% drop out, ending with a dense softmax activated layer over the 50 known writers [5].

Baseline Closed World Text Transcription Agent A baseline closed world agent for just text transcription was created for comparison to the open world

Conv2d	(16, 64, x)
BatchNorm2d	(16, 64, x)
LeakyReLU	(16, 64, x)
MaxPool2d	(16, 32, .5x)
Conv2d	(32, 32, .5x)
BatchNorm2d	(32, 32, .5x)
LeakyReLU	(32, 32, .5x)
MaxPool2d	(32, 16, .25x)
Conv2d	(48, 16, .25x)
BatchNorm2d	(48, 16, .25x)
LeakyReLU	(48, 16, .25x)
Dropout2d	(48, 16, .25x)
Conv2d	(48, 16, .25x)
BatchNorm2d	(48, 16, .25x)
LeakyReLU	(48, 16, .25x)
Dropout2d	(48, 16, .25x)
Conv2d	(64, 16, .25x)
BatchNorm2d	((64, 16, .25x)
LeakyReLU	(64, 16, .25x)
Conv2d	(80, 16, .25x)
BatchNorm2d	(80, 16, .25x)
LeakyReLU	(80, 16, .25x)
Flatten Interior	(1280, .25x)
Reshape	(.25x, b, 1280)
5 x bidirectional LSTM	(.25x, b, 512)
Linear	(.25x, b, 80)
LogSoftmax	(.25x, b, 80)

Table 6: Convolutional Recurrent Neural Network used for Handwriting Recognition in the Baseline Open World Agent. For experiments in which a CRNN embedding is used, the embedding is extracted at the double line. Note that 'x' is the input image width and 'b' is the batch size, which is only shown when it is not in the first position.

IAM Writer Identification Distribution for Basic Feature and Transcription Evaluation

Datasplit Type	Total Writers	**Total Writers in Split**			**Total Intersecting Between Pairs**		
		train	val	test	train & val	train & test	val &test
IAM Aachen	431	373	93	170	65	135	57
5-Fold CV	431	∼354	∼251	∼259	∼216	∼216	∼216

Table 7: The mean 5-fold cross validation experiment's distribution for IAM writer identification. The version of the IAM data is the Aachen version, which standardizes some of the character transcriptions and handles errors. RIMES was excluded from this due to all RIMES documents being treated as a single unknown writer. The used 5-fold cross validation indicates the approximate distribution of writers between each data split for a single round of training and evaluation of a fold.

agent described in the main text under novel conditions. This baseline agent produces text for each writing sample, based on what it knows from the 50 known writers in the training set, by applying log-softmax to the output of deep recurrent layers [8]. It serves to demonstrate limited utility only in a closed world, over-fit to known writers, with considerable degradation performance when exposed to novel conditions.

4 Basic Feature and Transcription Evaluation: Additional Protocol Information and Detailed Analysis

The data splits for the cross validation were obtained by first halving the unique writers into two equal groups of 216 each. Then, one half was randomly shuffled and split into 5 folds in a traditional 5-fold cross validation manner, stratified by the writer identifiers for the best representation of all 216 writers in each fold's samples. The other half was then further split into 5 groups of unique writers with no intersection. This second split ensures that for every fold, there is a set of novel writers in the test dataset. Each half's 5 folds were then aligned randomly such that typical 5-fold cross validation may occur. The training set, consisting of 4 folds, for every round of cross validation was then split in the exact same way, such that the validation set also had novel writers at evaluation time. This method of obtaining the 5-fold cross validation folds results in the approximate distribution of writers in training, validation, and testing sets as seen in Table 7. This is approximate, because due to the imbalanced number of samples per writer, where some writers had less than 5 samples, some folds had more writers than others.

4.1 Novel Characters in Transcription

In general, most HWR models will have some sort of "background" class to represent spurious marks or mistakes on the page. For the purposes of this experiment

we trained on a combined IAM and RIMES dataset in which the RIMES transcriptions were modified to include the characters that are not a part of the IAM dataset as background. RIMES and IAM were broken into folds such that Zipf's law gives us a variety of known unknown and unknown unknown characters in each fold, in the terminology of open world recognition. In terms of novelty, due to Zipf's law, novel characters unseen at training time occurred naturally in both the validation and testings sets for all folds. The addition of RIMES, and thus all of its French specific characters not included in the IAM dataset, included more characters whose labels were never known to the agent. However, some were included by design in the training set as unknown characters seen during training (known unknowns).

4.2 Novelty in Overall Image Appearance

In order to simulate novelty in the ODAI style recognition subtask, we augmented the IAM and RIMES datasets by randomly modifying the backgrounds of the images. The data was split into three different representation classes for training, Noise, which added Gaussian background noise as well as over the foreground, Antique, which adds a background similar to that of a historical document, and the Original White background from a clean document scan. The Original White background with black foreground is the unaltered background of each image found in IAM and RIMES, and is the typical ideal clean scan of handwritten documents. Additionally there are two known unknown classes that are seen at training time, Reflect_0, which is flipping the text image over the horizontal axis, and Blur, which adds Gaussian blur to the image. There is another augmentation only included in the validation and test sets, Reflect_1, which reflects the image over the vertical axis. Finally, the test set includes another novel image appearance class where the image has inverted color, the InvertColor class.

The Antique class used a set of free-to-use background images totaling in 16 different background images all accessed as of the data 2020-12-30. Nine of these images were taken from commons.wikimedia.org that were categorized as "old paper", "vintage paper", or as "parchment":

- El siglo de las tinieblas, o memorias de un inquisidor; novela histórica origina: https://commons.wikimedia.org/wiki/File:El_siglo_de_las_tinieblas,_o_memorias_de_un_inquisidor;_novela_hist%C3%B3rica_original_(1868)_(14590934239).jpg
- Old paper 1: `https://commons.wikimedia.org/wiki/File:Old_paper1.jpg`
- Old paper 3: `https://commons.wikimedia.org/wiki/File:Old_paper3.jpg`
- Old paper 4: `https://commons.wikimedia.org/wiki/File:Old_paper4.jpg`
- Old paper 6: `https://commons.wikimedia.org/wiki/File:Old_paper6.jpg`

- Old paper 7: https://commons.wikimedia.org/wiki/File:Old_paper7.jpg
- Vinatage Paper Texture: https://commons.wikimedia.org/wiki/File:Vintage_Paper_Texture_(9789792113).jpg
- Blank page, brown paper texture: https://commons.wikimedia.org/wiki/File:Blank_page,_brown_paper_texture_(14802136533).jpg
- Parchment 00: https://commons.wikimedia.org/wiki/File:Parchment.00.jpg

Besides these more authentic paper backgrounds, some artists' free-to-use interpretations of antique paper were also used:

- Pixabay empty brown canvas on Pexels: https://www.pexels.com/photo/abstract-ancient-antique-art-235985/
- HD paper texture by Imrooniel on Deviant Art: https://www.deviantart.com/imrooniel/art/HD-paper-texture-298160595
- 5 paper textures by MarshmellowHeaven that were stained with coffee and tea:
  - https://www.deviantart.com/marshmellowheaven/art/Texture-Paper-1-195235719
  - https://www.deviantart.com/marshmellowheaven/art/Texture-Paper-2-195236191
  - https://www.deviantart.com/marshmellowheaven/art/Texture-Paper-3-195236939
  - https://www.deviantart.com/marshmellowheaven/art/Texture-Paper-4-195237220
  - https://www.deviantart.com/marshmellowheaven/art/Texture-Paper-5-195237843

These background images were randomly selected for every handwriting line image that was chosen to be of the Antique class. All backgrounds were turned to grayscale, while remaining in RGB color space to match IAM and RIMES formats. Given the chosen background image, a cropping of that background that fit the size of the handwriting line image was selected and the handwriting line was laid over that cropped background. The exact procedure is available in the provided code, which will be publicly available upon publication. While the Large-Scale evaluation did not assess ODAI as the 5-fold CV experiments did in the main paper, it did use the same code with different backgrounds to assess how the baseline agents performed when the background changed.

5 Large-Scale 55K Test Evaluation: Additional Protocol Information and Detailed Analysis

The Mean HOG configuration of the baseline open world HWR agent was evaluated with 55,000 tests. We generate 5,500 tests based on experimental conditions. For each generated test, we create nine additional tests, re-ordering the test samples to average-out sample variations while retaining the same conditions of the

Training Set Mean Measures of 5-fold Cross Validation

Task	Multi-class Classif. with Novel Class		Binary Novelty Detection	
Model	NMI	Acc.	NMI	Acc.
Writer ID				
Mean HOG EVM	0.8198 ± 2.09e-3	0.8444 ± 8.58e-4	0.9557 ± 1.12e-3	0.9889 ± 3.38e-4
10-Mean HOG EVM	0.9871 ± 9.55e-4	0.9921 ± 4.37e-3	0.9586 ± 3.35e-3	0.9900 ± 9.38e-4
ResNet50 EVM	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
CRNN-PCA EVM	0.9996 ± 9.90e-5	0.9998 ± 4.07e-3	0.9990 ± 3.94e-4	0.9998 ± 6.18e-5
Appearances (ODAI)				
Mean HOG EVM	0.6611 ± 2.47e-3	0.8559 ± 1.46e-3	0.4817 ± 4.04e-3	0.7654 ± 2.78e-3
10-Mean HOG EVM	0.7124 ± 1.33e-3	0.8973 ± 5.74e-4	0.6076 ± 3.02e-3	0.8348 ± 1.74e-3
ResNet50 EVM	0.8327 ± 5.44e-6	0.8000 ± 4.61e-6	0.4326 ± 6.93e-6	0.6667 ± 6.40e-6
CRNN-PCA EVM	0.4455 ± 3.19e-3	0.6546 ± 5.27e-3	0.0840 ± 6.78e-3	0.3089 ± 1.31e-2
	Character Acc.	**Word Acc.**	**NMI**	**Acc.**
Transcription				
CRNN	0.9904 ± 5.83e-4	0.9660 ± 1.98e-3	0.9601 ± 3.17e-3	0.9913 ± 8.07e-4

Table 8: The mean 5-fold results with standard error for the train split of all three experiments. "NMI" stands for Normalized Mutual Information. All measures reported here are found after selecting the maximum probable class as predicted by the classifier after thresholding the maximum probability to determine if novel.

test. Tests were constructed and grouped by types of novelty. The tests were constructed to evaluate both single writing sample novelty detection and world change detection indicated by data distribution change from a non-novelty phase to a novelty phase of the test. In addition to establishing a foundation for novelty detection and characterization in HWR, in this evaluation, we establish some initial metrics for novelty difficulty, identifying factors impacting the performance of detecting novelty and transcribing handwritten text.

5.1 Protocol: Modified IAM Off-Line Handwriting Data.

The roughly 55,000 novel writing samples used in evaluation were constructed from modified samples of the IAM Offline Handwriting Dataset [4]. The training data will be publicly released after this paper's publication. A representative portion of the tests will be released as well.

Training and evaluation data, in the form of individual lines, was selected from IAM. Prior to training, lines were denoised, removing shadow boxes around the letters of each word. Features were then extracted from the clean lines of written text to capture writing characteristics including pen pressure, letter slant, word spacing and character size [2]. A distance matrix was formed between by the sum of absolute differences between each writer's mean style across all example words from each writer. The distance matrix served as a writer similarity measurement.

The training set was made up lines of text from 50 selected writers representing a subset of the writer style descriptor values, leaving one or two bins for each feature excluded for use in the novelty evaluation set. Lines of text did not

Validation Set Mean Measures of 5-fold Cross Validation

Task	Multi-class Classif. with Novel Class		Binary Novelty Detection	
Model	**NMI**	**Acc.**	**NMI**	**Acc.**
Writer ID				
Mean HOG EVM	0.7394 ± 9.28e-3	0.7123 ± 4.93e-3	0.7754 ± 1.66e-2	0.9265 ± 7.19e-3
10-Mean HOG EVM	0.6497 ± 7.90e-3	0.7852 ± 3.83e-3	0.3857 ± 8.17e-3	0.6246 ± 6.13e-3
ResNet50 EVM	0.6403 ± 8.56e-3	0.7876 ± 3.23e-3	0.3793 ± 6.18e-3	0.6126 ± 6.18e-3
CRNN-PCA EVM	0.6513 ± 7.95e-3	0.8074 ± 3.54e-3	0.3949 ± 6.96e-3	0.6266 ± 6.77e-3
Appearances (ODAI)				
Mean HOG EVM	0.5809 ± 2.54e-3	0.7886 ± 1.96e-3	0.3358 ± 3.86e-3	0.6464 ± 3.69e-3
10-Mean HOG EVM	0.4948 ± 4.14e-3	0.7525 ± 1.96e-3	0.2894 ± 2.75e-3	0.5799 ± 2.49e-3
ResNet50 EVM	0.0272 ± 1.18e-3	0.5097 ± 4.11e-4	0.0181 ± 7.61e-4	0.0989 ± 2.21e-3
CRNN-PCA EVM	0.0177 ± 1.87e-3	0.4315 ± 5.96e-3	0.0027 ± 1.42e-2	0.4848 ± 6.28e-3
	Character Acc.	**Word Acc.**	**NMI**	**Acc.**
Transcription				
CRNN	0.9516 ± 3.53e-3	0.8861 ± 2.61e-3	0.8787 ± 7.03e-3	0.9664 ± 2.44e-3

Table 9: The mean 5-fold results with standard error for the validation split of all three experiments. "NMI" stands for Normalized Mutual Information. All measures reported here are found after selecting the maximum probable class as predicted by the classifier after thresholding the maximum probability to determine if novel.

contain any additional effects, using a white background. Sample lines from six additional writers were chosen to compose an unknown writer training set. The set was supplemented with samples from the RIMES dataset and samples from the same 50 writers with background effects including salt and pepper noise, antique paper, and faded impressions of shaded boxes around the words in each line of text.

The evaluation set was made up of the remaining writers and writing sample manipulations to alter characteristics of both the writing style and the background. Letter style manipulations included thinning or widening, brightness, resizing and slant adjustments to each line of text. The background was composed from Creative Commons licensed images of textured paper. Pen manipulations were similarly constructed by merging in textures and colors, weighted by the pixel strength (*i.e.*, pen pressure).

The difficulty associated with each test is determined by the novelty type. The difficulty for novel writers and novel letter manipulations was determined by the ontological separation of four writing style features: pen pressure, letter slant, character size, and word spacing. Grouping novel writers with non-novel writers with similar styles is intended to make detection more difficult. The difficulty of novel pen and backgrounds was measured by the inverse intensity of the background (since the letters are black).

Most novel examples were constructed with a single type of novelty. Background and pen novelties were applied to sample text lines from the 50 known writers. The number of text lines per test varied based on availability of data targeting the specific novelty: 512, 768, or 1,024. In total, each test selected from

1,696 non-novel examples of the 50 known writers and approximately 50,000 novelties. Tests were composed of writing samples selected and organized by six independent discrete variables defining the experimental conditions of each test to explore the performance regime in novelty detection, resulting in 3,888 unique combinations. Using several subtypes (*e.g.*, different backgrounds) of novelties by type and difficulty, we constructed approximately 5,500 tests, each reordered nine times to average out sample variations.

Difficulty and novelty type (Table 10) affect writer prediction and transcription accuracy. Variables (Table 11) associated with distribution and placement of novelty in stream of data, such as introduction point, density of novel to non-novel samples and distribution type are varied to measure impact on novelty detection.

Novelty Type	**Count**
Background	17,662
Letter	11,868
Pen	11,289
Writer	8,427
No Novelty	1,696

Table 10: Number of writing samples for each type of novelty.

Independent Variables	**Values**
Mean Novelty Introduction Pt.	0.4, 0.5, 0.6, 0.7, 0.8, 0.9
Density of Novelty	6 different densities
Novelty Type	Writer, Letter, Background
Difficulty	Easy, Medium, Hard
Distribution Type	High (positive skew), Low (negative skew), Mid (normal), Flat (uniform)
Test Length	512, 768 and 1,024

Table 11: Independent Variables forming the experimental conditions of each Novelty Test

5.2 Supplemental Results for the Large-Scale 55K Evaluation

This section provides a more fine-grained analysis over the 55,000 tests presented to the closed world agents and the novelty detecting open world agent described in the main text of the paper. The agent configuration used for the open world

agent experiments utilizes HOG features for all style tasks. For this analysis, we present general novelty detection, text transcription and writer identification performance across all tests based on types of novelty.

Closed World Agents: Transcription and Writer Identification As expected, novelty negatively impacted the writer identification and sample transcription accuracy. Results are shown in Table 12. Mean Character Transcription Accuracy is reported as $1 - L(G_s, A_s)/\max(|G_s|, |A_s|)$ where L is Levenshtein Edit Distance, G_s is ground truth text for writing sample s, and A_s is the agent's predicted transcription for writing sample s, averaged over the ten variations of each test. Writer Identification Accuracy is reported as mean accuracy of the top-1 and top-3 predictions out of K+1 writers, where $K = 50$ for all tests, and the additional class is for novel writers.

Is Novel?	Mean Char. Acc.	Writer ID Top-3 Acc.	Writer ID Top-1 Acc.
False	0.85	0.99	0.99
True	0.47	0.40	0.24

Table 12: Baseline closed world agent mean character transcription accuracy, top-3 writer identification accuracy, and top-1 writer identification accuracy in response to non-novel and novel writing samples.

Open World Agent: Novel vs. Non-Novel Predictions Again as expected, novelty negatively impacted both text transcription and writer identification accuracy. However, the open world agent is significantly better at the text transcription task. Results are shown in Table 13. Transcription performance is reported as mean character accuracy computed using the ground truth and the agent provided transcriptions for all tests. Writer identification accuracy is reported as mean accuracy of the Top-1 and Top-3 predictions out of K+1 writers, where $K = 50$ for all tests.

Is Novel?	Mean Char. Acc.	Writer ID Top-3 Acc.	Writer ID Top-1 Acc.
False	0.82	0.942	0.719
True	0.62	0.479	0.220

Table 13: Baseline open world agent mean character transcription accuracy, top-3 writer identification accuracy, and top-1 writer identification accuracy in response to non-novel and novel writing samples.

Open World Agent: Novel Style Manipulations Style manipulations include manipulations to the characters. These manipulations had a measurable impact on writer identification performance. Results are shown in Table 14. Dilating the letters did not affect performance, down-weighting pen width as a major factor of a writer's style. More extreme character manipulations such as large slants and slants coupled with dilation were more easily detected as being novel, as expected. Inverting pixel values for written text did not adversely affect writer identification performance. The novelty detector did not equate letter inversion as novelty. Each novelty type was represented by 1,696 sample images.

Four different summary statistics are computed for the novel style manipulations. Novelty Detection Accuracy is mean accuracy of all of the detection decisions. Mean Character Transcription Accuracy is defined as

$$1 - L(G_s, A_s)/\max(|G_s|, |A_s|) \tag{2}$$

where L is Levenshtein Edit Distance, G_s is ground truth text for writing sample s, A_s is the agent's predicted transcription for writing sample s, averaged over the ten variations of each test. NMI represents normalized mutual information between the actual writer of the sample and the top-1 predicted writer. Writer Identification Accuracy is mean accuracy of the top-3 predictions out of $K+1$ writers, where $K = 50$ across all tests, and the additional class is for novel writers. These summary statistics are also used for the novel pens and novel backgrounds assessments, which are described below.

Novelty Type	**Novelty Detection Acc.**	**Mean Char. Acc.**	**NMI**	**Writer ID Acc.**
Dilate	0.99	0.70	0.01	0.57
Erode	0.79	0.77	0.35	0.68
Increase Size	0.99	0.33	0.01	0.02
Big Right Slant	0.79	0.62	0.21	0.09
Slant w/ Dilate	0.99	0.46	0.04	0.00
Big Left Slant	1.00	0.55	0.01	0.02
Small Slant	0.86	0.52	0.23	0.04
Inverted	0.33	0.71	0.79	0.94

Table 14: Novelty detection accuracy, mean character transcription accuracy, top-1 writer identification mean normalized mutual information, and top-3 writer identification accuracy given pen novelties grouped by novel style changes.

Open World Agent: Novel Pens Novel Pens include manipulations to written text, replacing the pixels with textures and colors, weighted by the intensity of the pen as described by pen pressure. Results are shown in Table 15. Pen

manipulations had minimal impact on writer identification performance. Each novelty type was represented by 1,696 sample images.

Novelty Type	Novelty Detection Acc.	Mean Char. Acc.	NMI	Writer ID Acc.
Blue Color	0.98	0.78	0.09	0.53
Brown Texture	0.74	0.75	0.43	0.79
Gold Texture	0.92	0.69	0.22	0.73
Rainbow	0.98	0.71	0.10	0.57
Red Color	0.98	0.70	0.10	0.53

Table 15: Novelty detection accuracy, mean character transcription accuracy, top-1 writer identification mean normalized mutual information, and top-3 writer identification accuracy grouped by novel pens.

Open World Agent: Novel Backgrounds Background manipulation had a more diverse impact on writer prediction performance than style manipulations. Results are shown in Table 16. NMI represents normalized mutual information between actual sample writer and top-1 predicted writer. Novel types of shadow boxes (from the uncleaned lines extracted from IAM) had the highest writer identification accuracy, perhaps due to similar associations made with these types of artificial irregularities in the training set. As with pen manipulations, more extreme manipulations resulted in higher detection accuracy. Increased texture interfered with the agent's ability to identify the writer.

Novelty Type	Novelty Detection Acc.	Mean Char. Acc.	NMI	Writer ID Acc.
Antique	0.40	0.80	0.47	0.39
Blue Fabric	0.98	0.40	0.10	0.42
Blue Color	0.98	0.69	0.01	0.54
Blue Wall	0.99	0.33	0.01	0.10
Brown Fabric	1.00	0.60	0.01	0.11
Crinked Paper	1.00	0.71	0.02	0.16
Gaussian Noise	1.00	0.14	0.00	0.16
Gold Wall	0.68	0.51	0.34	0.37
Rainbow Paper	0.98	0.25	0.12	0.54
Shadow Boxes	0.59	0.88	0.62	0.91

Table 16: Novelty detection, mean character transcription accuracy, top-1 writer identification mean normalized mutual information, and top-3 novel writer identification accuracy grouped by writing style.

Open World Agent: Writer Similarity in Novel Writer Discovery Each test is composed of sample writing from known and unknown writers. Here we find the minimum distance of an unknown writer across all known writers. We hypothesize that the greater the distance of writer style attributes of unknown writers with known writers, as captured in the ontological specification, the easier it is to detect a novel writer.

Surprisingly, the results did not show a strong correlation as expected. Table 17 shows the Pearson's correlation of each style attribute with detection and top-1 novel writer identification accuracy. We believe this due to two key factors: not enough variability in writing styles in the unknown population and the chosen set of attributes insufficiently capturing all of the essential characteristics of writing style. Pen pressure had the highest correlation of the four ontological specified factors. Collectively, a weak positive correlation did support the hypothesize. The proposed benchmark can be augmented with additional attributes, as the challenge problem evolves.

Style	**Novelty Det. Corr.**	**Writer ID Corr.**
Slant Angle	0.06	0.01
Skew Angle	0.02	0.04
Word Spacing	-0.02	-0.05
Pen Pressure	0.14	0.09
Character Size	0.03	0.04
Summed	0.12	0.18

Table 17: Novelty detection and novel writer identification correlation grouped by writing style.

Open World Agent: Factors in Novelty Detection A critical factor in the 55K tests is the density and location of novelty introduction — the switch between pre-novelty and post-novelty phases of the test given a stream of writing samples. This approach treats novelty as perceived world changing events rather than outliers, where confidence of novelty predictions increases as more novel examples are encountered in the data stream, increasing the body of evidence. With this approach, the level of false positives, those misidentified non-novel examples that fall in the pre-novelty phase of the test, can be substantially reduced. Fig. 4 shows the false positive count by the proportion of novelty. The variability and amount of false positives decreases as the proportion of novel samples to non-novel samples increases.

We conducted ANOVA to identify factors affecting the false positive rate (see Table 18). Along with the proportion of novelty, distribution type had a significant impact on the false positive rate. A positively skewed distribution, where novel samples densely occur at the start of the novelty phase of the test,

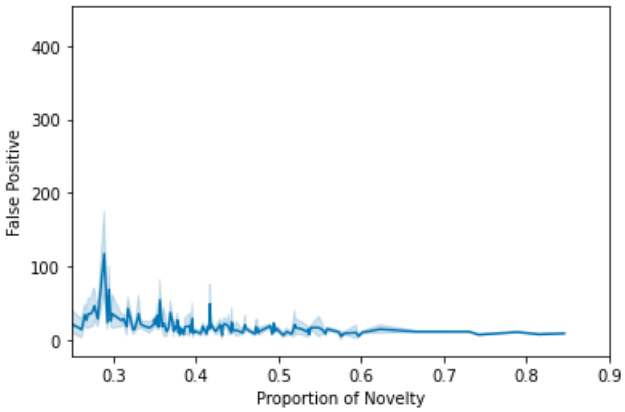

Fig. 4: False positive count by the proportion of novelty present.

is associated with lower false positive rate when compared to other distribution types such as a negatively skewed distribution. Novelty difficulty had a weak association to the false positive rate.

Factor	Sum of Squares	*df*	F	*p*
Distribution Type	1.058e+06	3	129.300	0.000
Level of Difficulty	7.456e+03	2	1.365	0.255
Location of Novelty	2.646e+06	1	969.301	0.000
Proportion of Novelty	5.848e+05	1	214.240	0.000
Residual	1.205e+00	44170		

Table 18: ANOVA analysis of statistical influence given several test generating independent variables identified in Table 11 on false positive rate.

References

1. Boult, T.E., Grabowicz, P.A., Prijatelj, D.S., Stern, R., Holder, L., Alspector, J., Jafarzadeh, M., Ahmad, T., Dhamija, A.R., Li, C., Cruz, S., Shrivastava, A., Vondrick, C., Scheirer, W.J.: A unifying framework for formal theories of novelty:framework, examples and discussion. In: AAAI (2021)
2. Joshi, P.M., Agarwal, A., Dhavale, A., Suryavanshi, R., Kodolikar, S.: Handwriting analysis for detection of personality traits using machine learning approach. International Journal of Computer Applications **130**, 40–45 (11 2015). https://doi.org/10.5120/ijca2015907189

3. Malemnganba, M.: ml-graphology. https://github.com/Malemm/ml-graphology (2018)
4. Marti, U.V., Bunke, H.: The IAM-database: an English sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition **5**(1), 39–46 (2002)
5. Reddy, T.: Offline handwriting recognition cnn. https://github.com/TejasReddy9/handwriting_cnn (2018)
6. Ross, D.A., Lim, J., Lin, R.S., Yang, M.H.: Incremental Learning for Robust Visual Tracking. International Journal of Computer Vision **77**(1), 125–141 (May 2008). https://doi.org/10.1007/s11263-007-0075-7, https://doi.org/10.1007/s11263-007-0075-7
7. Rudd, E.M., Jain, L.P., Scheirer, W.J., Boult, T.E.: The extreme value machine. IEEE T-PAMI **40**(3), 762–768 (2017)
8. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition (2015)
9. Vinciarelli, A., Luettin, J.: A new normalization technique for cursive handwritten words. Pattern Recognit. Lett. **22**(9), 1043–1050 (2001), http://dblp.uni-trier.de/db/journals/prl/prl22.html#VinciarelliL01