Twin Convolutional Neural Networks to Classify Writers using Handwriting Data

Andrew Lim, Danica Ommen
Iowa State University

Center for Statistics and Applications in Forensic Evidence

Link to video presentation

Project Rationale & Goals

- Offline handwriting samples rely primarily on visual properties for classification.
- Traditionally, handwriting identification is tasked by forensic examiners, though it is argued that using a set of visual criteria and human judgement is unreliable.
- Challenges in handwriting identification arise due to the minute details in a writer's style and even variations with individuals, which can be due to the environment, injury, or fatigue, as well as the lack of defined features in handwriting.
- In modeling real-world scenarios in which handwriting is found as part of recovered evidence it is highly unlikely that there exists a database of other handwriting samples of the suspects. As a result, we address writer classification through the common source problem rather than identifying the specific writer.

Common Source Problem

H_p: The two items of handwriting originate from the same writer.

H_d: The two items of handwriting originate from different writers.

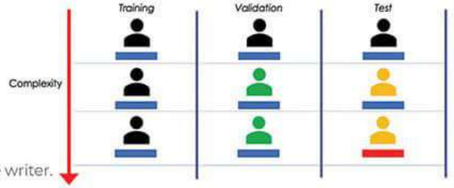


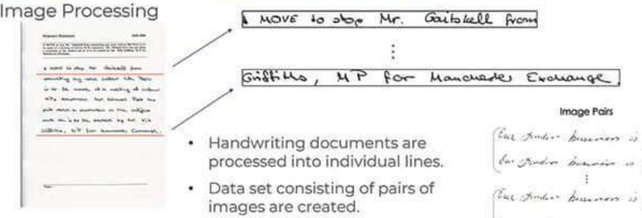
Figure 1: Writer Identification Complexity. The classification of writers becomes increasingly difficult for deep learning models when the pool of writers is mutually exclusive in each split of the data, and when images from a different source are classified.

Primary goals are to examine:

1. Writer diversification versus representation.
2. Preservation of handwriting structure versus image density.
3. Input size versus training size.
4. Writer identification complexity assessment using various test sets.

Image Processing & Sampling

Image Processing



- Handwriting documents are processed into individual lines.
- Data set consisting of pairs of images are created.

Figure 2: Paired Image Data Set. Pairs of images are formed as observations and given a response value of 1 if the two images are from the same writer, and a value of 0 if the two images are from different writers.

Sampling

- Data is naturally unbalanced with as there are many more pairwise comparisons with different writers than same writers.
- Number of sampled pairs is a very small fraction of the entire data set (Table 1). Methods to increase samples per writer.

1. Image augmentation.
2. Targeted sampling.
3. Subset of writers.

Data Set	Training Size	# Sampled Pairs	# Total Pairs	% Sampled
IAM Train	6161	10000	18,975,880	0.053%
IAM Train 40 Writers	1166	10000	679,195	1.47%
IAM Train Segmented 40 Writers	5028	10000	43,501,128	0.023%

Table 1: Number of total and sampled pairs in different data sets.

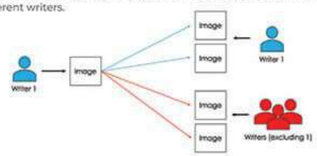


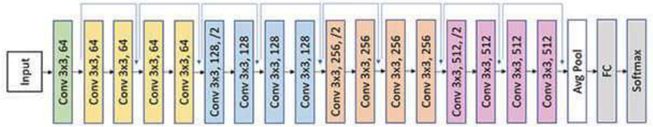
Figure 3: Targeted Sampling. An image from the data set is randomly sampled, then two images that are different but from the same writer and two images from the set of different writers are randomly chosen and paired with the first image.

Model Architecture

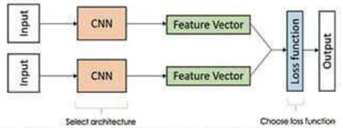
Custom Convolutional Neural Network (CNN) Architecture

Convolutional block consists of:
- Convolutional Layer
- Pooling Layer
- Batch Normalization Layer
- Dropout Layer

ResNet18 Architecture



Twin Convolutional Neural Network (TCNN) Architecture



- Selected CNN architectures include; Custom, ResNet50, ResNet18, VGG16, Xception.
- Model is optimized by the Euclidean loss function.

Results & Discussion

Training Size

Sampled a set of 10,000 training image pairs that were subsetted to smaller training sizes using nested random sampling.

Model	Training Size	Val Size	Val Acc	Val2 Acc
Custom	2000	2000	86.6%	76.55%
Custom	5000	2000	84.0%	75.6%
Custom	10000	2000	87.9%	85.5%

Table 2: Training Size on Validation Accuracy.

The results in Table 2 show that training on a larger training set does not necessitate drastic improvements in the prediction accuracy, unless we increase the number of training images significantly. However, due to memory constraints, we are limited in the amount of data we can pass into the model. Thus, the number of training pairs could not significantly exceed a training size of 10,000 pairs.

Figure 4: Nested Sampling. The same images are included in each subsequent larger data set as to reduce the variance due to a difference in training samples.

Input Size

Downsizing images reduces the number of features by distorting and reducing the pixel space. However, using larger images severely limits the number of passable inputs to the models.

Input Size	Training Size	Val Size	Val Acc	Val2 Acc
224 × 224	10000	2000	87.9%	84.15%
448 × 112	10000	2000	84.6%	70.4%
152 × 152	10000	2000	85.3%	77.96%
100 × 100	10000	2000	77.8%	73.7%

Table 3: Input Size on Validation Accuracy.

From the results in Table 3, we can see that for square inputs, the Val2 accuracy increased as the size of the images got larger. Though, for image sizes with a larger aspect ratio, the model performs worse since there is a lower pixel density leading to a lower Val2 accuracy.

224px x 224px 448px x 112px

Figure 5: Resized Images. Comparing a square input to an input that has larger aspect ratio, the writing in the square image is more distorted but yields a denser image array of handwriting features.

Segmented Images

Overall, for any training size, we see that using the resized data set as input to the Custom model, the number of correctly classified images in the Val2 data set is higher than using the segmented line data set as input.

Model	Training Size	Val Size	Val Acc	Val2 Acc
Custom	30000	10000	77.9%	69.1%
Custom	50000	10000	79.16%	73.2%
Custom	70000	10000	79.71%	71.84%

Table 4: Validation accuracy of segmented images.

Model	Training Size	Val Size	Val Acc	Val2 Acc
Custom	30000	10000	83.17%	78.64%
Custom	50000	10000	83.93%	79.66%
Custom	70000	10000	83.56%	79.28%

Table 5: Validation accuracy of resized images.

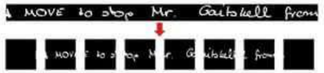


Figure 6: Segmented Images. Shows a line image being segmented into eight smaller images that contain parts of the line. Images that do not contain enough features, such as the first and last image in the partition are removed from the data set.

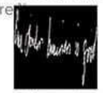
Model Selection

The Custom CNN with a much smaller number of layers than ResNet50 and Xception performed best, which indicates the high-level features extracted from deeper layers did not generalize well to other groups of writers.

Model	Training Size	Val Size	Val Acc	Val2 Acc
Custom	10000	2000	87.9%	85.5%
Xception	10000	2000	85.3%	79.65%
ResNet50	10000	2000	81.6%	78.4%
VGG16	10000	2000	78.4%	71.2%
ResNet18	10000	2000	72.64%	66.02%

Table 6: Validation accuracy of various models.

The Custom CNN embedded in the TCNN classified 82.8% of line images from the London Letter prompt and 50.2% of full documents from the CSAFE dataset correctly. Results are consistent up to the third level of complexity (Figure



Resized sample from London Letter data set.

Resized sample from CSAFE data set.

Conclusion

- Model is heavily constrained by limited computing resources.
- Using image augmentation, targeted sampling, and a subset of writers to increase writer representation in the samples gave poor results on the validation accuracy.
- Given the results, we found writer diversification and high image density to be necessary to properly train the chosen models. Whereas, both input size and training size were shown to have meaningful impacts on the classification accuracy.
- Evaluating the best model on images from a different data set gave similar results to evaluating on the same data set as the training data.
- Extending the model to evaluate entire documents did not perform well.
- Future development of GPUs will likely lead to increased GPU memory which can increase training and image size.

* List of references can be found in the full research article located at the Iowa State University digital repository.

This work was funded (or partially funded) by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements 70NANB15H176 and 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.