HWD: A NOVEL EVALUATION SCORE FOR STYLED HANDWRITTEN TEXT GENERATION

Vittorio Pippi, Fabio Quattrini, Silvia Cascianelli, and Rita Cucchiara

{name}.{surname}@unimore.it

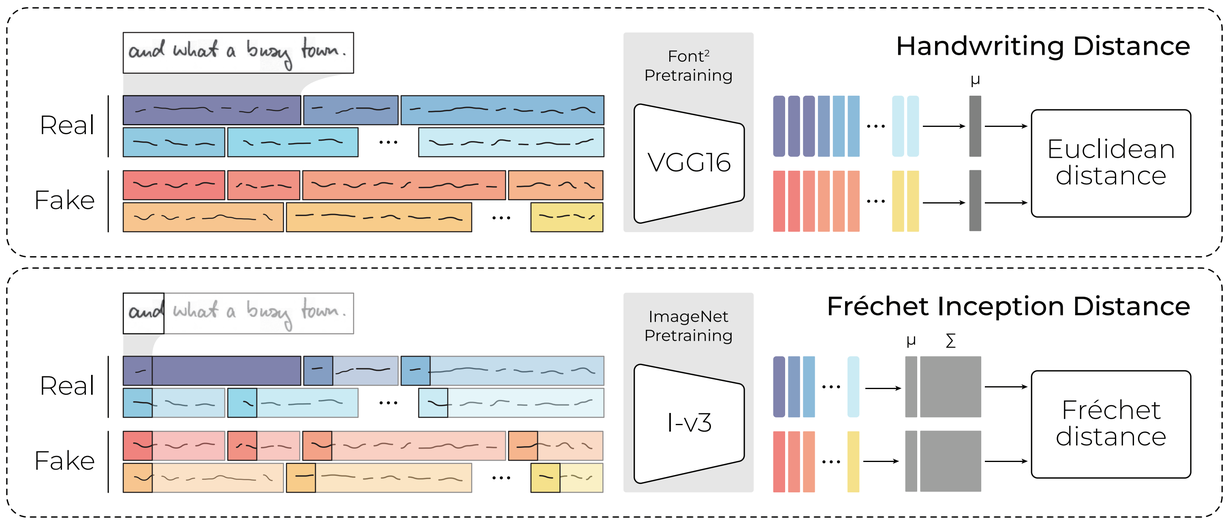
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

BMVC 2023

Overview

The Handwriting Distance (HWD) score is tailored for styled Handwritten Text Generation evaluation. In particular, it works in the feature space of a network specifically trained to extract handwriting style features from variable-lenght input images and exploits a perceptual distance to compare the subtle geometric features of handwriting.



Perception-Aware Feature Distance

The main idea behind distribution distance-based evaluation scores like the FID is to assess a generative model's performance based on its ability to produce images that align with the distribution of real ones.

In the context of Styled HTG, which focuses on handwriting and subtle geometric characteristics, a score that evaluates perceptual aspects is deemed more appropriate than one based on feature distribution distance. Therefore, we employ the Euclidean distance between the averaged feature vectors of the real and generated images in the style of the same writer:

$$Y_m = \frac{\sum_{i=1}^{N}\sum_{j=1}^{W_i} f(\mathbf{x}_{m,i})_j}{\sum_{i=1}^{N} W_i} \qquad \text{and} \qquad Y'_m = \frac{\sum_{i=1}^{N'}\sum_{j=1}^{W'_i} f(\mathbf{x}'_{m,i})_j}{\sum_{i=1}^{N'} W'_i}.$$

For the images in the style of writer m, the HWD is given by:

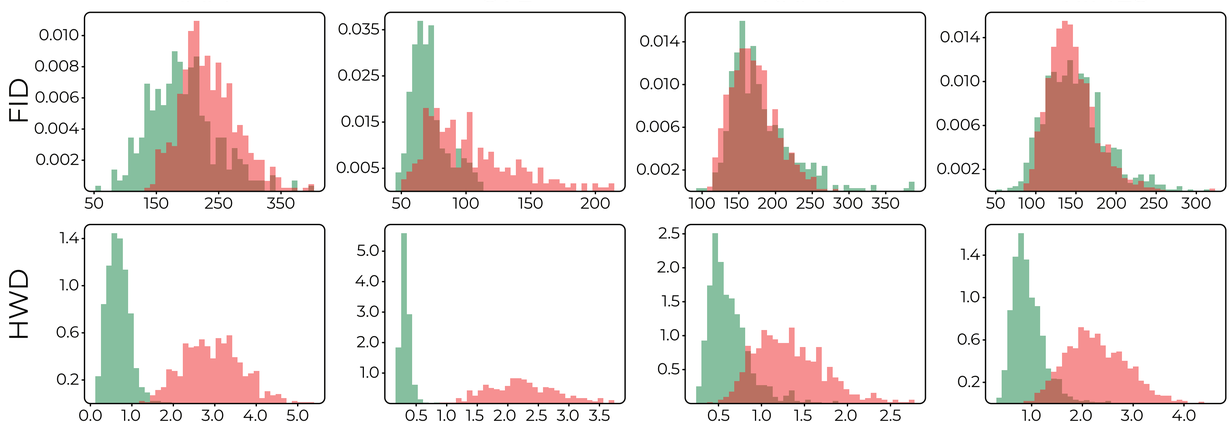
$$\mathrm{HWD}_m = \|Y_m - Y'_m\|_2.$$

Finally, the HWD on datasets containing images in the style of M different authors is obtained as

$$\mathrm{HWD} = \frac{1}{M}\sum_{m=i}^{M} \mathrm{HWD}_m.$$

Sensitivity to the Handwriting

We compare the HWD score against the FID score in the variant proposed in [2], which is the common approach adopted in Styled HTG.

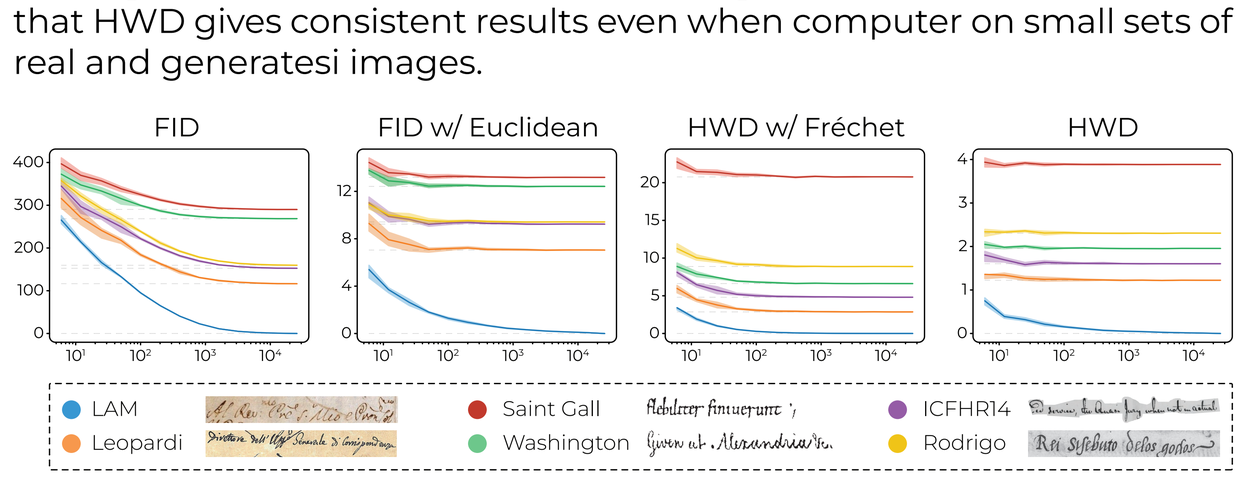


From the distribution of the HWD and FID scores when applied on same-author (green) or different-author (red) subsets emerges that with HWD It Is easier to separate same-author pairs from different-author ones.

	Language	Samples	Authors	Avg. Samples per author	FID		HWD	
					Overlap	EER	Overlap	EER
Norhand	Norwegian	21939	12	1828.25	4.2	4.2	0.0	0.0
BanglaWriting	Bengali	17265	212	81.44	11.6	5.6	6.1	2.9
CVL	English/German	13473	310	43.46	24.7	12.5	0.0	0.0
IAM	English	13353	657	20.32	27.1	13.6	0.7	0.3
KHATT	Arabic	11427	838	13.64	40.3	21.6	12.0	5.9
RIMES	French	12111	1500	8.07	39.1	20.8	7.0	3.3

Sensitivity to the Number of Samples

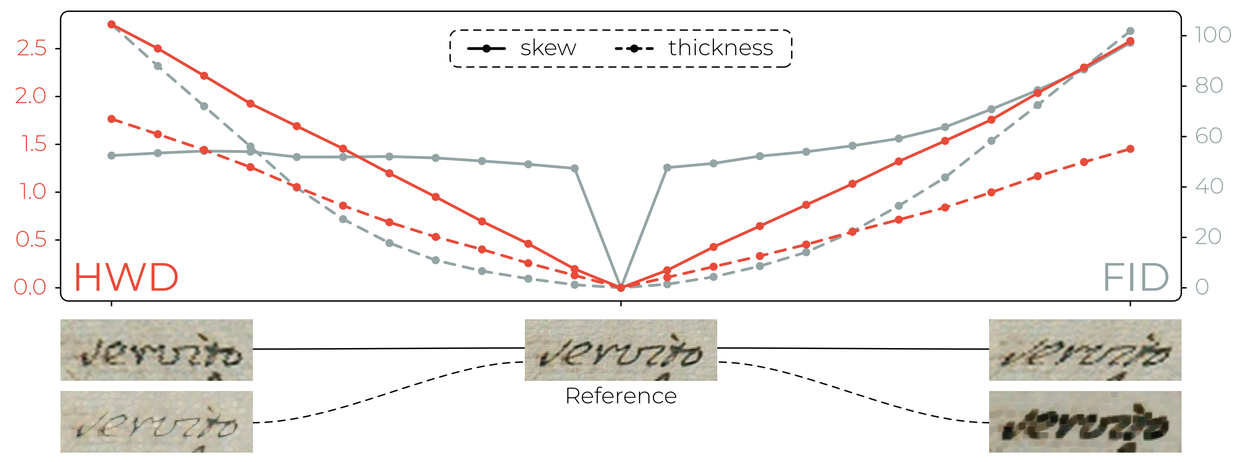
As argued by many publications, the FID exhibits a strong bias towards the number of samples. To assess the numerical stability of HWD, we consider the large single-author LAM dataset and compute the values of the HWD and FID on images from LAM against variably-sized subsets of images from different datasets including LAM itself. The results show that HWD gives consistent results even when computer on small sets of real and generatesi images.



Sensitivity to the Visual Appearance

To assess the sensitivity to handwriting-related visual aspects, we compare the FID and HWD between reference images and increasingly altered ones, taken from the LAM dataset. In particular, the considered alterations entail shear, erosion, and dilation to simulate handwriting slant and strokes thickness.

The HWD exhibits a more linear and interpretable behavior with respect to increasingly-severe alterations.



Ablation Analysis

We analyze the four key components of our proposed scoring system: the backbone model, the pretraining dataset, the input image portion, and the distance metric. This analysis is conducted using the IAM dataset, and the results highlight the importance of the backbone model, particularly VGG16 pretrained on Font², in influencing the score's quality.

Backbone	Pretraining Dataset	Image Portion	Distance	IAM	
				Overlap	EER
Inception-v3	ImageNet	Beginning	Fréchet	27.1	13.6
Inception-v3	ImageNet	Beginning	Euclidean	29.6	14.5
Inception-v3	ImageNet	Whole	Fréchet	24.0	11.6
Inception-v3	ImageNet	Whole	Euclidean	8.5	3.9
Inception-v3	Font²	Beginning	Fréchet	18.8	9.3
Inception-v3	Font²	Beginning	Euclidean	11.3	4.8
Inception-v3	Font²	Whole	Fréchet	19.0	9.1
Inception-v3	Font²	Whole	Euclidean	7.2	3.3
VGG16	ImageNet	Beginning	Fréchet	3.2	1.6
VGG16	ImageNet	Beginning	Euclidean	26.2	13.0
VGG16	ImageNet	Whole	Fréchet	2.8	1.2
VGG16	ImageNet	Whole	Euclidean	6.2	2.9
VGG16	Font²	Beginning	Fréchet	3.4	1.7
VGG16	Font²	Beginning	Euclidean	16.5	8.2
VGG16	Font²	Whole	Fréchet	3.5	1.6
VGG16	Font²	Whole	Euclidean	**0.7**	**0.3**

References

1. Zhang R., Isola P., Efros A., Shechtman E., and Wang O. *The unreasonable effectiveness of deep features as a perceptual metric.* In CVPR, 2018.
2. Kang L., Riba P., Wang Y., Rusiñol M., Fornés A., and Villegas M. *GANwriting: Content-Conditioned Generation of Styled Handwritten Word Images.* In ECCV, 2020.