

Investigating the Effect of using Synthetic and Semi-synthetic Images for Historical LULEĂ **Document Font Classification**

Konstantina Nikolaidou¹, Richa Upadhyay¹, Mathias Seuret², Marcus Liwicki¹

¹ Machine Learning Group, Luleå University of Technology, Sweden, firstname.lastname@ltu.se ² Pattern Recognition Lab, Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany, firstname.lastname@fau.de

Highlights

Use font classification dataset of early printed books [21]

Pre-process it to get image crops

Create baseline dataset of 10K crops

Investigate SOTA pre-trained CNN architectures on baseline and on baseline with additional 60K:

- semi-synthetic images from DocCreator [10]
- synthetic images from OpenGAN [4]
- real data samples

	Dataset	ResNet50	DenseNet	EfficientNet
	Baseline	95.78 ± 0.55	96.91 ± 0.41	96.79 ± 0.10
	+DocCreator	97.93 ± 0.21	98.30 ± 0.69	97.85 ± 0.24
	+GAN	95.41 ± 0.33	96.81 ± 0.43	96.40 ± 0.42
	+Real	96.57 ± 0.23	96.89 ± 0.54	97.16 ± 0.12

OF TECHNOLOGY

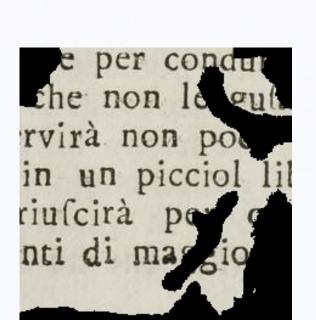
Classification accuracy on the competition test set.

Semi-synthetic data

one per condurre che non le gust: rvirà non poco in un picciol lil riuscirà per ogi nti di maggior o

one per condurre che non le gust: rvirà non poco in un picciol lil riuscirà per og nti di maggior d

one per condurre che non le gust rvirà non poco in un picciol lil riuscirà per og nti di maggior



Original

one per condurre che non le gull: rvirà non poco in un picciol lil riuscirà per ogi nti di maggior c

Character degradation

Bleed-through

one per condurre che non le gust: rvirà non poco in un picciol lil riuscirà per ogi nti di maggior c

> Phantom character

Blur

one per condurre che non le gust: rvirà non poco in un picciol lil iuscirà per og nti di maggior

Shadow

Holes

instance per degradation (6 total) using random parameter values in DocCreator

We created 1



DenseNet accuracy starting with the 10K baseline samples and with gradual addition of DocCreator samples.

Synthetic data

We used the default parameters of OpenGAN

REAL

GAN

GAN

Gotico Antiqua

Secunda est. vi no pu maniadalistaminatione est. qua sic poterit certifici tud ninus ninus ninus iiti aque prenalere frin

maqua. Ipse aut pfl 100min vister winn vister manu digito terat et o na 3 dands dands dands d lignate. non colori. qi minuminin ailima ilima i rolum distimuler. Si winin vistremillio millio

Hebrew



REAL

Rotunda

Co Inin O Incyclaru ciome que era bien a cuyo puecbo est lupiessen q es cola ecida y autorizada euerendissima:pa exemplo del zeloy for vent mavor

noth lock lock lock anda be hihon bet ie besit bur besit bo mind am hind anor nour location loca ciboa besiiboa bet ie berichus beriche mind am. hind amer

Fraktur

gangen oco Calcor noch die übrigen a eschichte, den Sofi et und Evagrius ül Publikum zu entsche 21 4

poedeine tie deloe Hot. Tota fet. 188 ee db.ed dh e Coinclot fofn fot

Key Takeaway

Domain knowledge provided by DocCreator degradations surpasses fully synthetic data created by a GAN and even contributes to the performance more than real data.

References

[21] Seuret et al. (2019). Dataset of Pages from Early Printed Books with Multiple Font Groups. Proceedings of the 5th International Workshop on Historical Document Imaging and Processing.

[10] Journet, et al. (2017). DocCreator: A New Software for Creating Synthetic Ground-Truthed Document Images. J. Imaging, 3, 62. [4] Ditria et al. (2020). OpenGAN: Open Set Generative Adversarial Networks. ACCV.