Digital Humanities: Pre-Processing Historical Text Images for Generative Adversarial Neural Networks Using Linear and Non-Linear Transforms

Gregory Heyworth Christopher Bruinsma Syed Shihan
gheywort,cbruinsm,sshihan @u.rochester.edu

May 3, 2023

Abstract

Multi-Spectral Imagery provides a more detailed view of physically damaged historical manuscripts. The goal of Linear and Non-Linear transforms as well as industry standard techniques to provide clarity on text that may be damaged chemically, physically, or may have eroded the parchment or vellum the text is written on. The goal of current research is to find the best transforms best removes the damage on these historical manuscripts. The contribution of this research is two-fold: the use of novel and standard procedures to reveal the under-text in image and the quantifying of the usability of these images for use in GANs which are at present, absent from the field.

Introduction

Manuscripts are pivotal to digital humanities and often contain pieces of literature that have been unseen to the eyes of history for hundreds of years. These Manuscripts may not also be unseen, but also contain texts that help to complete our understanding of a particular body of work. These have included works such as new-found works by Apuleius or complete versions of the Roman code of laws known as the Giaus.

Our combination of standard and non-standard algorithms will pre-process our images and generate image pairs. The standard algorithms include Principle Component Analysis (PCA), Minimum Noise Fraction (MNF), Independent Component Analysis (ICA), and Spectral Angle Mapper (SAM), performed in NV5 Global Spatial's ENVI software. The non-standard algorithms used are Kernel Principle Component Analysis (KPCA) and Texture Co-occurrence. Additionally, we will implement in Python, Keith Knox's Blur and Divide algorithm, which involves dividing an image that has undergone a median filter pass by a second image that has undergone a Gaussian filter pass. This research aims to determine which analysis methods and algorithms are most effective in producing image pairs that provide the greatest amount of contrast between over and under texts, and similarity to an ideal example of a text with good contrast and legible writing. Qualitative analysis will assess how well the processes segment the images to human observers. This will help identify the most suitable for training a Generative Adversarial Network (GAN) that can be used to denoise future manuscripts. This study contributes novel methods to the field, as well as the use of GANs.

Based on current research's understanding of linear and non-linear methods for image processing in addition the non-linear nature multi-spectral data gathered in the digital humanities, we expect the non-linear methods to perform better for image pre-processing than linear methods. We expect the way to handle the non-linear illumination from inks captured through various chemical reagent degradation will be separable by non-linear filters as well as wavelength specific techniques such as SAM. While the research of Giacometti et al. shows the robustness of PCA and ICA on documents that have been physically or chemically degraded current research argues that because of the non-linearity of some feature spaces in multi-spectral imagining using a non-linear feature space could produce even more robust results.[3]

Principle Component Analysis

Principle component analysis is a common tool in digital humanities and has key uses for textual recovery especially in Multi-spectral images. PCA is a form a preprocessing which can order data in terms of its most and least useful components. PCA helps us to find which pieces of our data items are more useful and then in essence find ways to more completely separate data. For image processing and indeed the digital humanities this can be used to better separate components of an image.

For textual recovery, this can be used to more clearly delineate text from the background of the image such as the parchment of vellum it is written on. This is used for correlation and is an orthogonal linear transform.

$$Z = XS^{-1}$$

Kernel PCA

Kernel PCA uses a kernel function to project a dataset into a higher dimensional feature space, where it is separable. Unlike PCA, Kernel PCA is a non-linear method which allows for the classification of data whose decision boundaries are described by non-linear functions. The idea behind this process is translating the data to a higher dimension space in which the decision boundary becomes separable. Through doing so, Kernel PCA extends the functionality of a normal PCA by allowing for the separation of non-linear data using convolution kernels. As a whole, this process is used for novelty detection and image de-noising. The following describes the process for calculating the symmetrical RBF kernel matrix.[7]

$$\kappa(x_i, x_j) = exp(-\gamma \|x_i - x_j\|_2^2)$$

ICA

Independent component analysis or ICA is used when there are multiple input signals coming in at once, and the goal is then to uncover their signals separately from the whole. The goal of this as opposed to PCA is for Independence between variables. ICA is used as a linear reduction method. This can be defined simply as follows:[9]

$$p(x, y) = p(x)p(y)$$

MNF

Maximum noise fraction (MNF) is composed of two separate principal components analysis rotations; the first rotation conducts the process of "noise whitening" by using the principal components of the noise covariance matrix to de-corellate/re-scale the noise in the data, the second rotation uses the principal components derived from the original image data after they have been noise-whitened to divide the data space into two parts and separate the noise from the data. This process is used for hyper-spectral imagery denoising, much like PCA. In MNF the noise fraction of the "$i^{th}band$" from Green et. al is defined as:[5]

$$Var\{N_i(x)\}Var\{Z_i(x)\}$$

This is then transformed linearly as follows:[5]

$$Y_i(x) = a_i^T Z(x) \text{ for each } i \text{ in the bands.}$$

Spectral Angle Mapping

Spectral Angle Mapping is a way to threshold an image based on the spectral features of an image. This can be done to classify components of an image base on their wavelengths in nano-meters. The can be a good way of separating components of an image based on their reluctance of certain wavelengths of light or to find the components of an image which are known to fluoresce at a certain rate under those same wavelengths.[1]

$$\alpha = cos^-1 \frac{\Sigma XY}{\Sigma(X^2)\Sigma(Y)^2}$$

Texture Co-Occurrence

Texture Co-Occurrence is an algorithm that is used a spatial measure of the number of times pairs of pixel values occur in an image. Our images are grey-scale and this transform and this algorithm works to create a grey-scale co-occurrence matrix in order to measure: Contrast, Correlation, Energy, and Homogeneity. This can be a guiding measure because under-text, over-text, and chemical reagents often have strictly different grey-scale magnitude pixels. This is calculated and built into what is known as a GLCM or Grey Level Co-Occurrence Matrix. Each of the components of Contrast, Correlation, Energy, and Homogeneity are defined as follows:[6, 4]

1. Energy = $\Sigma_{i,j} p(i,j)^2$
2. Contrast = $\Sigma_{n=0}^{N_{g-1}} n^2 \left\{ \Sigma_{i=1}^{N_g} \Sigma_{j=1}^{N_g} p(i,j) \right\}$
3. Correlation = $\frac{\Sigma_i \Sigma_j (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
4. Homogeneity = $\Sigma_{i,j} \frac{1}{1-(i-j)^2} p(i,j)$

Software

ENVI Software from L3Harris

ENVI is an industry leading software used for image processing provided by L3Harris. Its application in digital humanities is its ability to combine and process multi-spectral imagery. Often times data can become more or less separable based on which wavelengths are being processed through a process known as triage. Our motivation behind using ENVI is its ability to bridge across the powerful features of IDL to python and vice versa allowing us to develop and deploy our own code in ENVI to take advantage of its workflow.

Python: Algorithms

Current research will be using Python to write both the Blur and Divide and KPCA Algorithms. In addition all of our metrics which include RMF-contrast (RMF-Score) and Frećet Inception Distance (FID Score).

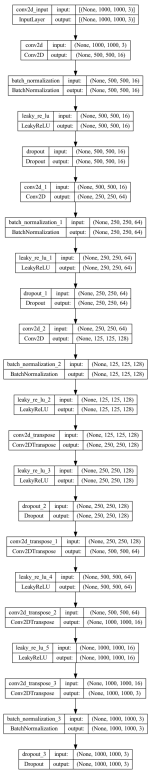
Python: TensorFlow.Keras

Our GAN will be built based on the technologies of Tensorflow and will be run on a NVDIA GeForce GPU's. Our Generator Model will be based on a Tensoflow's DCGAN taking in images cropping them to a 1600 x 1600 image and then performing a series 2D convolutions at various sizes from 16 $\rightarrow$ 128 in size followed by Batch normalization's and PreLU layers on the decent.[2] On the ascent the model uses 128 $\rightarrow$ 16 2D transposes followed by batch normalization and PreLU layers.

Our logic behind Parametric Rectified Linear Units is their ability to be used for incredibly robust neural networks that are less prone to over fitting. Given our incredibly limited quantity of data this was our method for reducing this risk.

The Generator Model has been trained on 46 images in pairs. Our generator is then fit to the transformed images and is then trained alongside the the Discriminator in a secondary training step which allows the GAN to remove the darker pixels from the page and reveal the lighter text below.

We propose a model has 315,000 trainable parameters and 422 un-trainable parameters that looks as follows:



Procedures

Spectral Triage

The selection of images taken in the Lazarus process include three types of images. (1) Images that use reflected light, (2)images that are taken with transmitted light, and (3)images taken using fluorescence that are filtered through a wheel of filters that allow only a certain wavelength of fluoresced light through. All of which has been provided by MegaVision.

In order to find the best spectral bands [bands] one first needs to go through the 40-50 image bands and select the images with the most contrast in and between the other images. This is done to take what is called a spectral subset. This spectral subset then needs to be calibrated.

Using ENVI, a Spectralon key from the photo is taken using a small pure-white cube in each image using the region of interest tool. These calibrated images are then saved and will be used to perform all of the transforms. The statistics for each transform can be performed using both a smaller region of interest in the image that usually contains the best contrast between over-text and under-text and some of the damage consistent with these often Palimpsest documents. This is not avaible to the python scripts.

Performing the Transforms : ENVI

Performing the transforms of PCA,ICA,MNF,and Texture Co-Occurrence using ENVI is a simple procedure and produces several images that are usually qualitatively examined for their ability to be read. Using the metrics of FID and Contrast. In addition for MNF, PCA, and ICA ENVI provides the associated Eigenvalues. The Eigenvalue metrics from these transforms provide the quantity of information available across each of the bands.

Performing the Transforms : Python

In python, one simply needs to select the same *.tif* files that are included in the ENVI spectral subset for the same other transforms. This is how Blur and Divide and Kernel PCA are run. Kernel PCA in our case has been used with both RBF, Cosine, and Sigmoid Kernels after this has been done both the eigenvalues and histograms are saved. For Blur and Divide each of the images are then saved as tif files and can then be examined in the same vein as the images from ENVI.

Collecting Data : Python

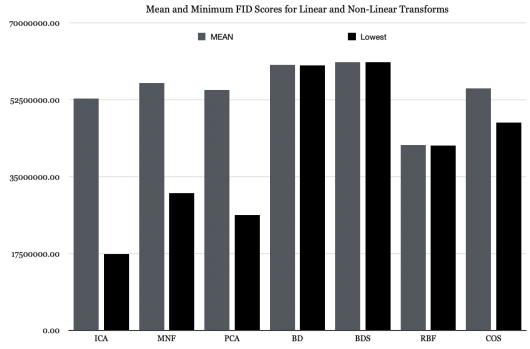
Once the images have been recorded from ENVI and Python, they then have their histograms, contrast, and FID scores taken. This data is then saved to associated .csv and .png files. All of our code is located here: https://github.com/KodakC41/LAZGAN

Results

For the transforms, our quantification is twofold. First we examined the mean FID scores and second we examined mean RMF contrast. For our GAN generated images which for now were based purely on the Blur and Divide category for reasons that will be described later are similarly measured based on their FID scores compared to their transformed counterparts. This differs from the Transformed images which are measured based on their distance from an image that has been transformed and then manually edited in Adobe Suite's Lightroom and Photoshop. Using Kernel PCA and comparing it to PCA as a proxy to compare linear and non-linear transforms will be done through eigenvalue comparisons across the Our GAN itself will be measured using the metrics Mean Squared Error loss[MSE] and Fréchet Inception Distance [FID]. A noteworthy piece of data that has not been included are histograms. The information from each tells current research which images have the highest concentration of valuable information and which images have the highest quantities of clipped or unavailable data.

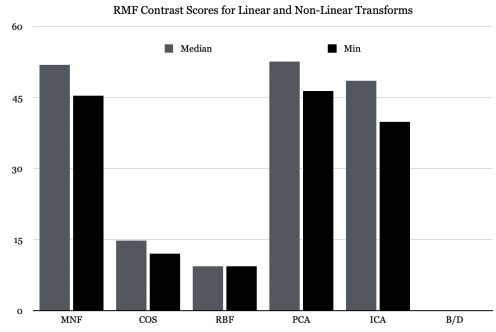
FID Results

These results are based on this image which has been transformed using ICA in ENVI and then edited in Adobe Lightroom to have higher contrast and inverted colors:

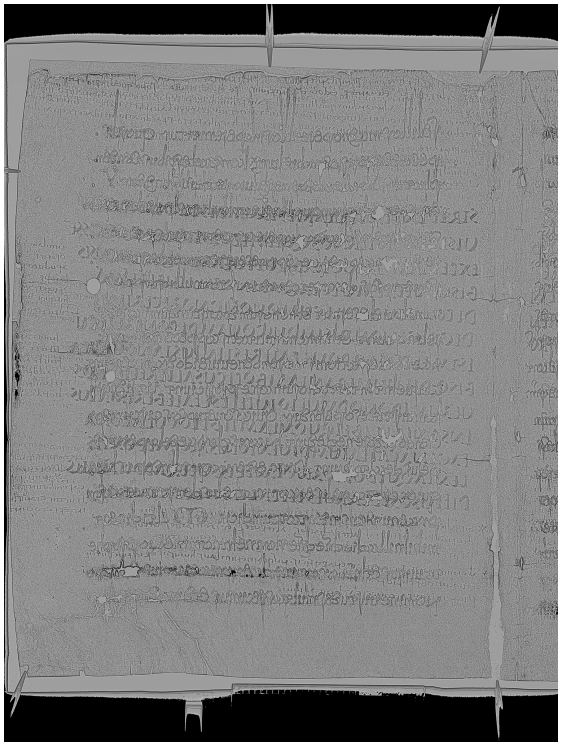


The benefit of calculating the FID score against this *pseudo*-ground truth image is the each image under review shows not only the under-text more clearly but also sharpens the text that may be damaged on the top and left of the image as well as the Palimpsest data towards the center of the images.

RMF Contrast



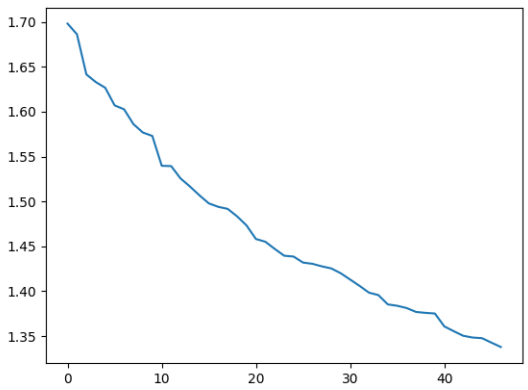
Here we can see that most of the the linear transforms generate equivalent amounts of contrast information for a viewer. Radial Basis Function KPCA and Cosine KPCA both provide much less contrast and Blur and Divide provides the least. Intuitively, this would mean that the linear transforms relay more information via their contrast. Contrary to this, Blur and Divide with its minimal contrast best compacts the information. This is shown through the clarity of the output images which compresses the over and under texts. As a second benefit, Blur and Divide works well to remove the chemical reagents from the lower third of this page which lends itself well to our proposed GAN. It was due to the efficacy of Blur and Divide that current research saw a way to use the pairs of images generated through our batch-blur-and-divide python script as training data four our textual recovery GAN or LAZ-GAN.



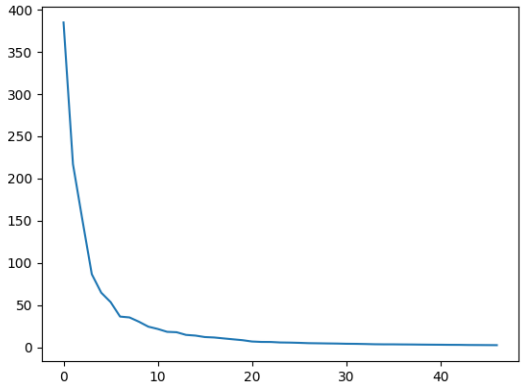
Eigenvalue Comparisons KPCA and PCA

In the digital humanities one method of comparison between different transforms is the Eigenvalue graphs between band rotations. For example, using eight spectral bands for PCA will yield a graph that uses each band as a principal component. The eigenvalue graphs tells the viewer how much information is elucidated across each band. This generally decreases across the bands as the information overlap increases. We can compare the efficacy of our experiments across full spectrum KPCA and PCA using these graphs to see how much information we can gain by performing each.

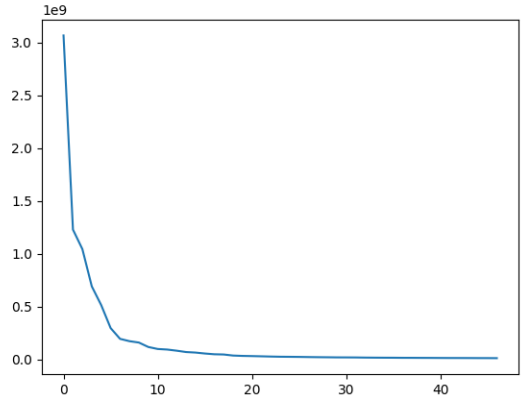
RBF-Kernel PCA



Cosine-Kernel PCA



PCA



From these plots what is shown that the Radial-Basis-Function Kernel provides the most linear decrease in information across all 47 bands. This is a helpful demonstration of how separable Kernel PCA can make our data. That being said, the amount of information as a factor of separability is incredibly narrow as opposed to both the Cosine-Kernel PCA and Linear PCA. Both provide minimal separability across spectral bands beyond the fourth or fifth band that being said the quantity of information available is much higher. However, it may be that similarly to blur and divide, RBF Kernel PCA reduces the dimensions of these images to a degree that they are easily separable visually. That being said, current research let go of Kernel PCA both from the findings of its lack of information across bands and at the introductory advice received on KPCA.

GAN Statistics and Example Results

We run this model three different ways, with a pre-compiled and fit generator model with Five Epochs across the entire data-set, with pre-compiled and fit discriminator and without any pre-fitting. Then we train

the both the discriminator and the generator in batches of 10.

Statistics - w/o Any Pre-Compilation

Epochs	Gen MSE	Disc MSE	Mean FID[8]
5	2023.35	0.032	1.15E6
10	2030	0.0037	1.16E6
15	2038.43	1.0402	1.17E6
20	2045.82	0.0118	1.20E6
50	1982.12	3.9202	1.14E6
100	1948.18	0.00	1.08E6

Statistics - Pre-Compiled Generator

Epochs	Gen MSE	Disc MSE	Mean FID[8]
5	2011.8	0.5818	1.14E6
10	2007.61	0.0161	1.20E6
15	2006.01	27.41	1.92E6
20	2060.02	8.407	1.18E6
50	1992.56	0.0136	1.14E6
100	1974.88	0.00	1.11E6

Statistics - Pre-Compiled Discriminator

Epochs	Gen MSE	Disc MSE	Mean FID[8]
5	2015.92	0.169	1.15E7
10	2042.32	8.7051	1.17E6
15	2047.60	47.4537	1.18E6
20	2060.03	0.00	1.19E6
50	1975.21	0.00	1.12E6
100	1966.14	0.00	1.10E6

*Note Regardless of the results in this table, the images produced by this extra discriminator step were almost entirely unintelligible. They had very little data or where entirely clipped to black until the Epoch range reaches 50+.

Example Output

Discussion

Our trials of linear and traditional methods against non-linear methods for creating training data sets have yielded noteworthy results that have implications for future machine learning based attempts at Palimpsest data segmentation. These results also provide some counter-intuitive insights into some of the difference between what the right levels of contrast are for a machine learning model verses a human observer. We note that for linear methods across the 47-spectral bands in the data set the contrast scores being high and FID distance being far from the ideal image may be worse for a human viewer than images with low contrast scores. This is evidenced by the relative increase in quality to an expert between over and under text when blur and divide is used verses PCA, KPCA, MNF or ICA.

In addition, based on these preliminary tests the linear methods remain a better choice for image segmentation than non-linear KPCA. While possibly advantageous as a layer in future GAN models for this purpose as a singular transform is inconclusive in its results. While more difficult to perform, KPCA has results in the RMF contrast and FID score range that approach those of Blur and Divide. That being said, another reason for hesitancy is the available information across bands in the Eigenvalue decomposition. RBF-KPCA provides less information across all of its bands than PCA despite KPCA's linear decay against the decay seen in PCA across the same data. Given the former and latter the non-linear method of KPCA remains a subject of interest but at present cannot be recommended against other traditional methods.

That being said, further trials will be needed to further evaluate which transforms are advantageous for training a GAN. With MSE loss in the 2000s and the relative image quality being sub-par we note that a decrease

in contrast associated with training the model singularly on images from Blur and Divide may not provide the best results. It is noteworthy that the GAN itself seems to be capable of segmenting the text from the page but not at all adept at segmenting palimpsest texts. Our recommendation to future research is to add in a Kernel PCA layer into the GAN model itself to more easily transform the data in a non-linear model.

Finally we recommend that future research investigate the results if texture co-occurrence and spectral angle mapping more thoroughly as methods for generating paired data set images. While both were performed their performance was not evaluated leaving this as a charge for future research.

Acknowledgements

We thank Dr. Heyworth for access to the Lazarus Labratory a wealth of data and advice. We thank Dr. Hanneken for advice on Kernel PCA and the introductory implementation. We thank Brian Griglak for advice on L3's IDL and ENVI. Finally we thank Jenny Bloom for additional IDL advice and access to the NV5 Geospatial support team. Without their advice this project would not have been possible.

References

[1] "de Carvalho Júnior, Osmar & Meneses, Paulo. (2000). Spectral Correlation Mapper (SCM): An Improvement on the Spectral Angle Mapper (SAM)."

[2] "Deep Convolutional Generative Adversarial Network :   Tensorflow Core. TensorFlow. (n.d.). Retrieved April 30, 2023, from https://www.tensorflow.org/tutorials/generative/dcgan"

[3] "Giacometti A., Campagnolo A., MacDonald L., Mahony S., Robson S., Weyrich T., Terras M., Gibson A., The value of critical destruction: Evaluating multi-spectral image processing methods for the analysis of primary historical texts, Digital Scholarship in the Humanities, Volume 32, Issue 1, April 2017, Pages 101–122, https://doi.org/10.1093/llc/fqv036"

[4] "Gebejes, A., Huertas, R., Tremeau, A., Tomic, I., Biswas, P. R., Fraza, C., & Hauta-Kasari, M. (2016). Texture characterization by grey-level co-occurrence matrix from a perceptual approach. Color and Imaging Conference, 24(1), 271–277. https://doi.org/10.2352/issn.2169-2629.2017.32.271"

[5] "Green, A. A., Berman, M., Switzer, P., & Craig, M. D. (1988). A transformation for ordering multispectral data in terms of image quality with implications for noise removal. IEEE Transactions on Geoscience and Remote Sensing, 26(1), 65–74. https://doi.org/10.1109/36.3001"

[6] "Harlick, R. M., Shanmugam, K., & Dinsten, I. (n.d.). (PDF) textural features for image classification - researchgate. Research Gate. Retrieved April 29, 2023, from https://www.researchgate.net/publication/302341151_Textural_Features_for_Image_Classification"

[7] "Raschka, S. (2014, September 14). Kernel tricks and nonlinear dimensionality reduction via RBF Kernel PCA. Sebastian Raschka, PhD. Retrieved April 28, 2023, from https://sebastianraschka.com /Articles/2014_kernel_pca.html"

[8] "Shi, J., Zu, N., Xu, Y., Bui, T., Dernoncour, F., &; Xu, C. (2021, June 24). Learning by Planning: Language-Guided Global Image Editing."

[9] "Talebi, S. (2023, April 1). Independent Component Analysis (ICA). Medium. Retrieved April 28, 2023, from https://towardsdatascience.com/independent-component-analysis-ica-a3eba0ccec35"