Machine Learning investigation of GaN dislocations in an unsupervised focus.

Albes Kotzai[1,*] & Dr Ben Hourahine[1]

1.Department of Physics, University of Strathclyde, Glasgow G4 0NG , Scotland, UK
*albes.Koxhaj-or-Kotzai.2015@uni.strath.ac.uk

1 Introduction

- In recent years, GaN has revolutionized solid state lighting, resulting in the creation of a multi billion dollar industry exploiting the development of LEDs composed of GaN and related materials. However, GaN does have a high density of structural defects. These are detrimental to the performance of semiconductor devices and as a result more research is being carried out to improve the semiconductor material used in modern LEDs.[1]

- Analysing the high number of dislocations visible in ECC (ElectronChanneling Contrast) images is challenging when many are taken each day. A more automatic approach is needed and can be handled with Machine Learning algorithms.

2 Machine Learning (ML)

- The name machine learning was coined in 1959 by Arthur Samuel in his paper “Some Studies in Machine Learning Using the Game of Checkers” (Fig.1)[2] Machine learning explores the capacity of a computer to learn from experience, i.e. to modify its processing on the basis of newly acquired information.

- In Layman terms, it is a form of statistical algorithms that predicts datapoints by finding patterns through their distances.

- A common form of ML is regression, for example used in Excel’s trendline method to fit curves to data and predict behaviour with respect to other variables in general. (Fig.2)

- Three main categories of ML are named after the way data is handled, being: *Supervised* if the data is labeled, *Unsupervised* if its not and *Reinforcement Learning* if its a reward system.

Fig.1: Arthur L. Samuel

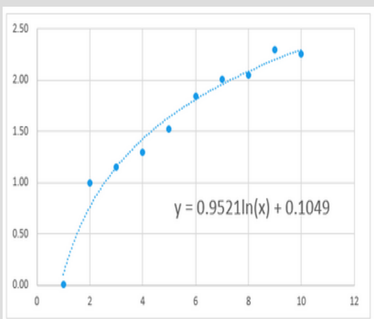

Fig.2: Example of Excel’s regression method. In essence, the y function is our ML model for predicting the future data points!

3 PCA

- Images can be composed of a very high number of pixels and even ML models can be slow in their training or investigation. This is where Principal Component Analysis (PCA) comes into play by lowering the dimensionality of the data used.[2-4]

- As shown on Fig.3 it identifies the highest variances in the data and uses those principal component vectors to summarise the data-set rather thanusing all of the data.

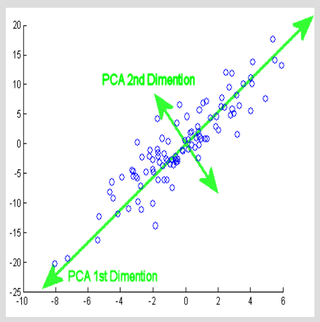

Fig.3: PCA example

4 Eigenfaces

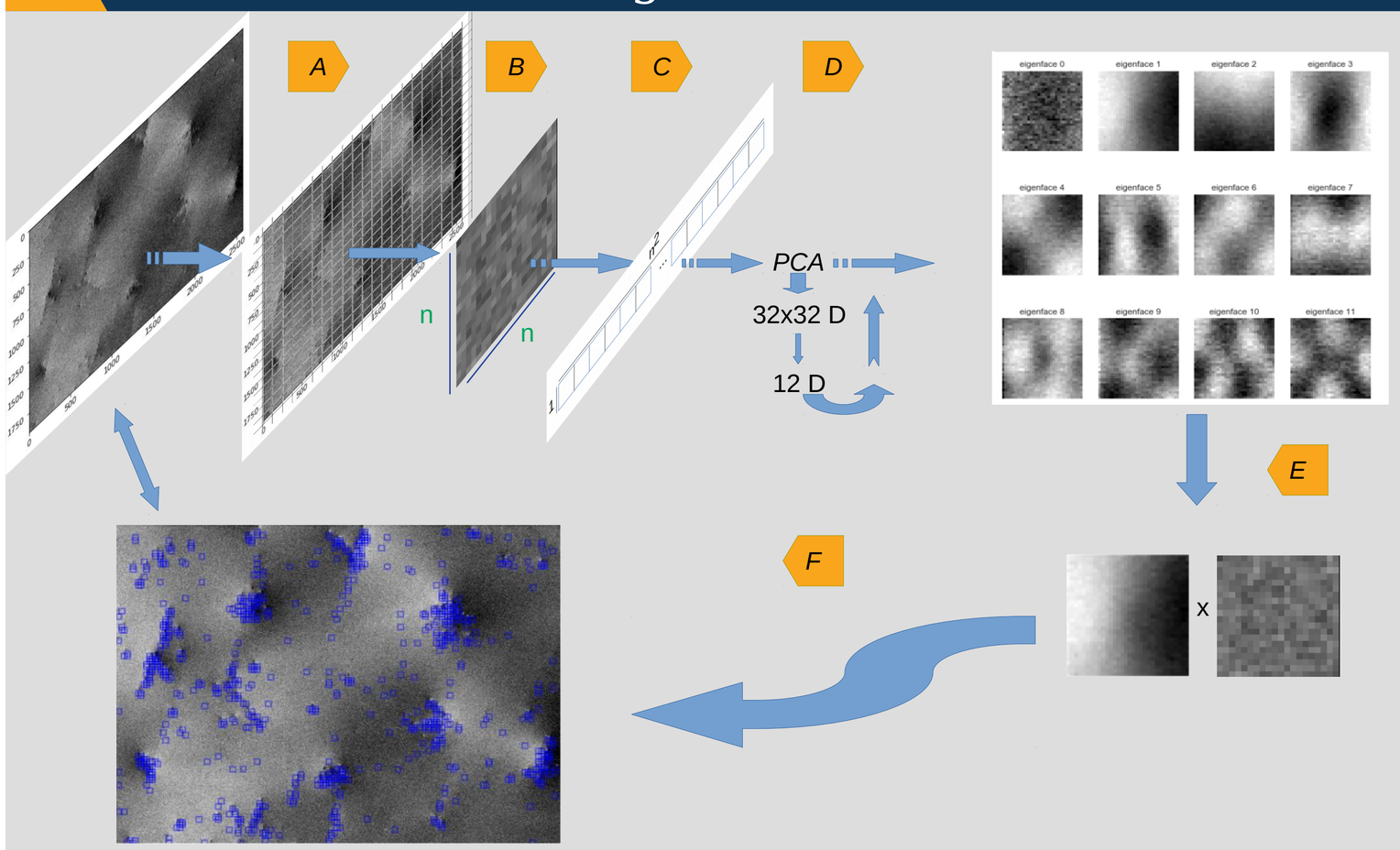

Fig.4: Step by step process of the approach used briefly with Eigenfaces >***A)*** Test ECC Image split in grid of nxn cuts with random choice of 50-80 patches (n=32) for unsupervised training ***B)*** Sample patch ***C)*** Unfolding the patch to a single vector ***D)*** Using PCA to lower its dimensionality from 32x32 pixels to only using 12 components from it to build our Eigenvectors and re-transform them back to Eigenfaces showing the variance picked ***E)*** Scalar product between the Eigenfaces and the ***all*** the patches of the test image not just 50-80. ***F)*** Result by applying a threshold.

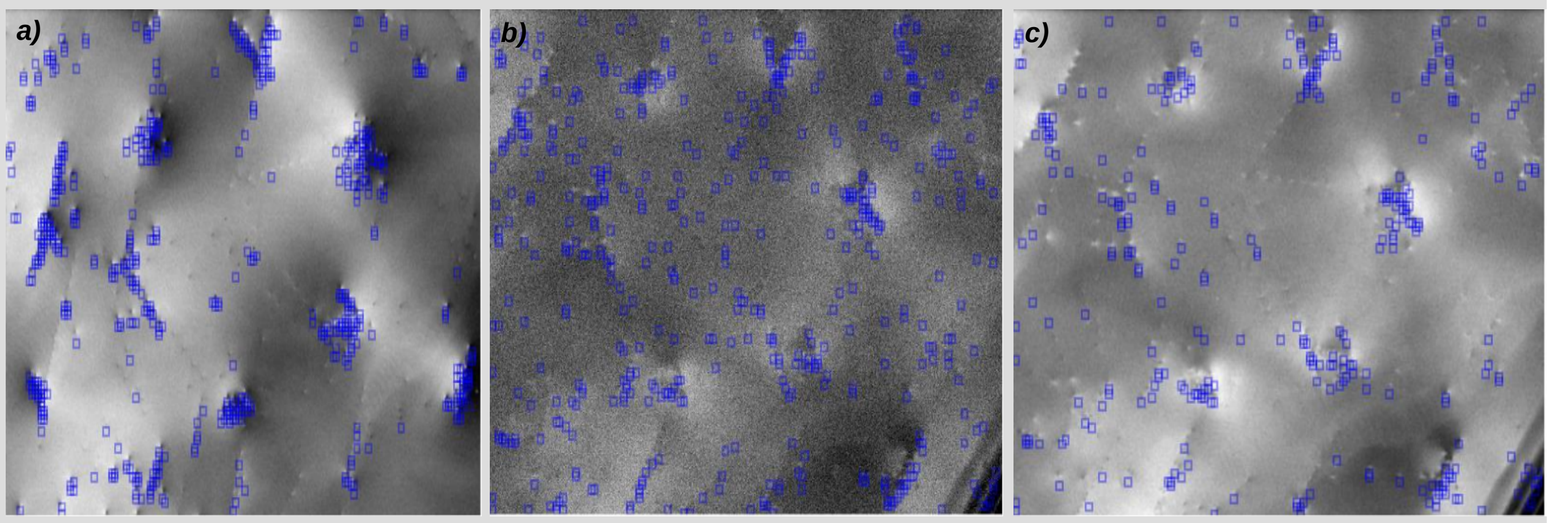

Fig.5 a) Test Image Smoothed. b&c) ***New*** and noisy image used with the same trained model. Results show clearly in unfiltered and filtered case *generalization*.

- As it can be seen from Fig.4 the approach of using unlabeled grid of data-sets with patches composed of 32x32 pixels in steps A) to F) shows remarkable results albeit using only 50-80 patches from an image with almost 5 million pixels. Furthermore, the approach takes a matter of second to be trained and a matter of minutes for the scan and window by window comparison of the scalar product to identify possible dislocations. In Fig.5 the generalization is shown as well with a new set of patches from another image not trained and showing that the PCA has done its job right enough for the Eigenfaces to be more unbiased. However, being unsupervised method without labels makes cross validation more difficult and in general false positives more abundant.

5 Conclusion

- Machine Learning algorithms are revolutionary [4]. Supervised methods, the majority of the programs currently used, are too costly and time consuming and usually require the need of expert labeling and immense databases. In this project a new approach is tried to improve upon known problems in unsupervised approaches and as a possible preprocessing step for supervised algorithms. The result is an unsupervised model based mostly on PCA, showing very promising results from only minutes of training time.

6 References

[1] https://journals.aps.org/rmp/issues/87/4
[2] Sarah Guido, Andreas Müller, *"Introduction to Machine Learning with Python-A Guide for Data Scientists", (2016)*
[3] Manohar Swamynathan, *“Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python “, (2016)*
[4] Nishant Shukla, “Machine Learning with TensorFlow”, (2017)