HAAG Weekly Report

Nikita Angarski – 3D Modeling

Week 9

Time-Log

- What did you do this week?
  - Implemented the preliminary testing suite comparing the PCA approach to our pycpd to the vanilla cpd, and vs. the low rank eigenvalue. Some interesting results, but nothing too spectacular yet.
  - Hosted and managed AV for the second seminar for the programs team.
- What are you going to do next week
  - Continue trying to implement testing using the included pycpd methods, but also open to exploring new methods to verify the new method accuracy.
  - Implement Bayesian cpd to test against the existing and new methods
- Blockers, things you want to flag, problems, etc.
  - Would need some more ideas on how to tune/reorganize the code to get it performing better than the vanilla method.

Abstracts:

Application of Principal Components

Analysis and Gaussian Mixture Models to

Printer Identification

https://www.cerias.purdue.edu/assets/pdf/bibtex_archive/nip04-ali.pdf

Abstract: Printer identification based on a printed document has

many desirable forensic applications. In the electropho-

tographic process (EP) quasiperiodic banding artifacts can

be used as an effective intrinsic signature. However, in text only document analysis, the absence of large midtone areas makes it difficult to capture suitable signals for banding detection. Frequency domain analysis based on the projection signals of individual characters does not provide enough resolution for proper printer identification. Advanced pattern recognition techniques and knowledge about the print mechanism can help us to device an appropriate method to detect these signatures. We can get reliable intrinsic signatures from multiple projections to build a classifier to identify the printer. Projections from individual characters can be viewed as a high dimensional data set. In order to create a highly effective pattern recognition tool, this high dimensional projection data has to be represented in a low dimensional space. The dimension reduction can be performed by some well known pattern recognition techniques. Then a classifier can be built based on the reduced dimension data set. A popular choice is the Gaussian Mixture Model where each printer can be represented by a Gaussian distribution. The distributions of all the printers help us to determine the mixing coefficient for the projection from an unknown printer. Finally, the decision making algorithm can vote for the correct printer. In this paper we will describe different classification algorithms to identify an unknown printer. We will present the experiments based on several different EP printers in our

printer bank. The classification results based on different

classifiers will be compared

Summary: **This research paper investigates methods for identifying printers based on the intrinsic signatures found in printed documents, specifically focusing on text-only documents. The authors explore the use of Principal Component Analysis (PCA) to reduce the dimensionality of data extracted from individual characters, followed by classification using Gaussian Mixture Models (GMM) and Classification and Regression Trees (CART). The core idea is that subtle, consistent imperfections (banding artifacts) in the printing process can serve as a unique fingerprint for each printer. The paper details the application of these techniques and presents experimental results demonstrating their potential for accurate printer identification in forensic applications.**

What did you do and prove it

Link to Seminar youtube: https://youtu.be/ejFVIb3-ig8

Link to testing suite commit: https://github.com/Nikitos1865/pycpd-Porto/commit/8bae205417cce6e37420d84e08ce25621bb9ae71