EAFS 2018
8th EUROPEAN ACADEMY OF FORENSIC SCIENCE CONFERENCE
LYON CONVENTION CENTER - FRANCE

2020 THE FORENSIC ODYSSEY
AUGUST 27th-31st 2018

Open Source Methods for Pattern and Impression Evidence Comparison

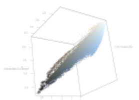
Mark J. Lancaster, Collaborator with the Center for Statistics and Applications in Forensic Evidence (CSAFE),
Iowa State University

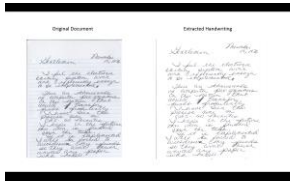
Open Source Methods for Pattern and Impression Evidence Comparison

Introduction

In previously presented research at EAFS 2015 and ICFIS 2017, it was illustrated by the author that open-source/nonproprietary image processing and machine learning techniques could be used to extract handwriting from documents and find useful features for assessing similarity of handwriting between two different documents, such as signatures.

For example, ink could be digitally separated from paper that had ruled lines. This can be handled through the use of a k-means clustering algorithm applied to the color space. Usually a small number of clusters is sufficient to obtain the handwriting component.




Original Document
Extracted Handwriting

Materials and Methods

One of the harder aspects of handwriting analysis is finding data that can be used without restriction. Researchers at the University of Terhan [1] have generously supplied the community with the **UTSig** database of Persian signatures that allows other researchers to use the image files. Here are genuine signature examples from the first three writers:

This dataset captures variability within the writer, as there are 27 signatures per writer (115 writers total). Nine signatures were captured on 3 different days. This data was captured in in a lossless image format (TIFF).

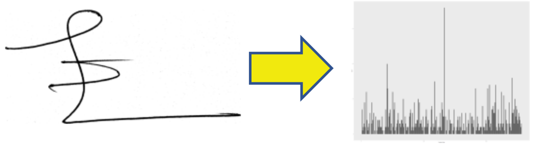
The VLFeat [2] open source computer vision library needs to have data in PGM format, and there is a nice (free) conversion tool XnConvert [3] that can process files in a batch mode. It can utilize the existing file structure of the data as well.

For this small analysis, the first 10 writers in **UTSig** are used to build the codebook information for use in a Latent Dirichlet Allocation (LDA) model. From these 270 images, feature

(Continued)

vectors that are 128-dimensional are extracted using SIFT [4], and the resulting dataset is then used to create a codebook with 256 words. This process is known as vector quantization (VQ). Once this codebook is built, one can find the appropriate code word for a new vector ***v*** by finding the centroid closest to it, and then assigning it the corresponding code word value.

This allows a signature image to be converted into a histogram, with the frequency of a code word corresponding to the number of times a feature was extracted.



Writer 11
Signature 1

Histogram
Encoding

EAFS 2018

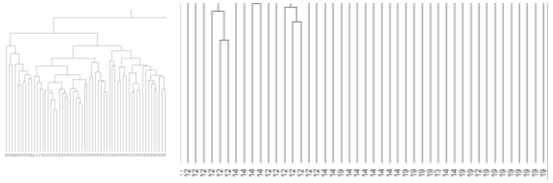
8TH EUROPEAN ACADEMY OF FORENSIC SCIENCE CONFERENCE
LYON CONVENTION CENTER – FRANCE

2020 THE FORENSIC ODYSSEY
AUGUST 27th-31st 2018

Open Source Methods for Pattern and Impression Evidence Comparison

(Continued)

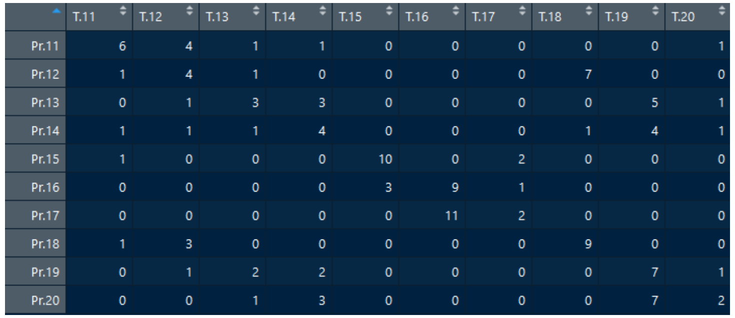
Note that even with black-and-white images, there is a signal that can be found using the SIFT features. As an example, if one uses Euclidean distance as a measure of similarity, one can build a dendrogram to see how well signatures cluster by writer:



Since we have a codeword representation of signatures, we can also use bag-of-words topic models such as LDA.

To use a topic model, one needs to build a document-term matrix and specify the number of latent topics that should be modeled. The processes to build the appropriate data structure and LDA model are all part of the **topicmodels** package for R. By using an LDA model for Writers 11 – 20 (14 in training, 13 in testing per writer) and 5 latent topics, one is able to obtain the following confusion matrix:

Results



	T.11	T.12	T.13	T.14	T.15	T.16	T.17	T.18	T.19	T.20
Pr.11	6	4	1	1	0	0	0	0	0	1
Pr.12	1	4	1	0	0	0	0	7	0	0
Pr.13	0	1	3	3	0	0	0	0	5	1
Pr.14	1	1	1	4	0	0	0	1	4	1
Pr.15	1	0	0	0	10	0	2	0	0	0
Pr.16	0	0	0	0	3	9	1	0	0	0
Pr.17	0	0	0	0	0	11	2	0	0	0
Pr.18	1	3	0	0	0	0	0	9	0	0
Pr.19	0	1	2	2	0	0	0	0	7	1
Pr.20	0	0	1	3	0	0	0	0	7	2

The accuracy based on these settings is 43%.

Discussion and Conclusion

Since LDA is a topic model, one can do further investigation into what components make a signature. This matches the natural language processing (NLP) idea that documents are made of a mixture of topics. For a difficult signature set (fewer branch and end points) compared to other languages, the application of topic models beyond their original use in NLP shows potential for pattern/impression evidence.

Bibliography

[1] A. Soleimani, K. Fouladi, B. N. Araabi. UTSig: A Persian offline signature dataset, *IET Biometrics*, 2017, 6(1), pp. 1-8.

[2] VLFeat: http://www.vlfeat.org

[3] XnConvert: https://www.xnview.com

[4] Lowe, D.G. Distinctive image features from scale-invariant keypoints *International Journal of Computer Vision*, 2004, 60(2), pp. 91-110.

[5] D. M. Blei, A. Y. Ng, M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 2003, 3, pp. 993-1022.

Contact Information

Dr. Mark J. Lancaster
Email:
markjlancaster@gmail.com
Phone:
+1 (859) 333 6383