Sonderforschungsbereich 950
Manuscript Cultures in Asia, Africa and Europe

CENTRE FOR THE STUDY OF MANUSCRIPT CULTURES

Writer Identification for Historical Manuscripts: Analysis and Optimisation of a Classifier as an Easy-to-Use Tool for Scholars from the Humanities

Hussein Mohammed [1], Volker Märgner [1][2], H. Siegfried Stiehl [1][3]

[1] SFB 950 / Centre for the Study of Manuscript Cultures (CSMC), Universität Hamburg, Hamburg, Germany, Email: hussein.adnan.mohammed@uni-hamburg.de
[2] Technische Universität Braunschweig, Braunschweig, Germany, Email: maergner@ifn.ing.tu-bs.de
[3] Department of Informatics, Universität Hamburg, Hamburg, Germany, Email: stiehl@informatik.uni-hamburg.de

Abstract

We analyse a state-of-the-art method w.r.t. common degradation types in historical manuscripts using images from the virtual manuscript library of Switzerland. Furthermore, we show that, by optimising a key parameter, we can enhance the performance of the method and significantly outperform the winner method of the Historical-WI competition. Finally, we demonstrate the practicality of our implementation through intuitively comprehensible results for direct by scholars from the humanities.

Motivation

So far, no thorough analysis is available in the literature concerning the impact of the degradation typically found in the digitized manuscripts on classification . Furthermore, the currently proposed methods are beyond the reach of the scholars from the humanities; either because of the impracticality of the method itself in real-world applications, or because of the lack of an easy-to-use implementation.

Contributions

- A thorough analysis of a state-of-the-art method w.r.t. two common degradation in historical manuscripts
- Parameter optimisation for enhancing the performance of the method on historical manuscripts
- A practical implementation of the method with an easy-to-use user interface and intuitive results presentation

State-of-the-Art Classifier:

Normalised Local NBNN Classifier [1]

It is a classifier for offline, text-independent, and segmentation-free writer identification based on the Local Naïve Bayes Nearest-Neighbour (Local NBNN) classifier [2], which is reformulated mathematically as follows:

$$Dist^{c}_{local} = \sum_{i=1}^{n} \left[\left(\parallel d_i - \phi(\mathrm{NN}_c(d_i)) \parallel^2 - \parallel d_i - \mathrm{N}_{k+1}(d_i) \parallel^2 \right) \right],$$

$$\hat{C} = \underset{C}{\operatorname{argmin}} \left(Dist^{c}_{local} \right),$$

where

$$\phi(\mathrm{NN}_c(d_i)) = \begin{cases} \mathrm{NN}_c(d_i) & \text{if } \mathrm{NN}_c(d_i) \leq \mathrm{N}_{k+1}(d_i) \\ \mathrm{N}_{k+1}(d_i) & \text{if } \mathrm{NN}_c(d_i) > \mathrm{N}_{k+1}(d_i), \end{cases}$$

This method takes into consideration the particularity of handwriting patterns by adding a constraint to prevent the matching of irrelevant keypoints:

$$|Ort_{kpt1} - Ort_{kpt2}| \leq T_r$$

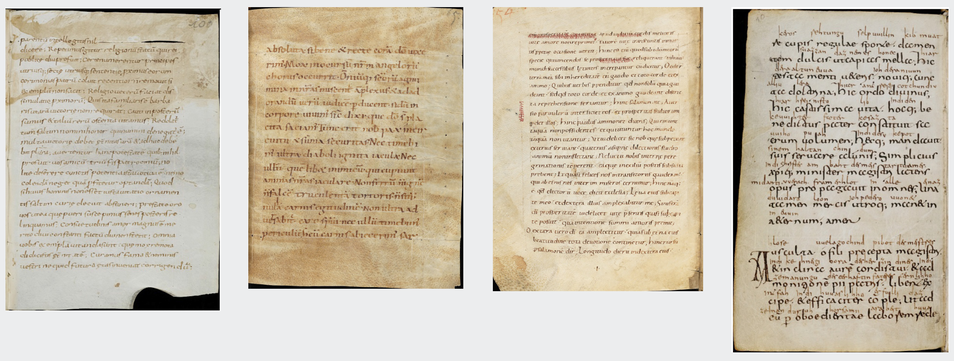
Furthermore, a normalization factor is used to cope with the problem of scarce and unbalanced data:

$$\hat{C} = \underset{C}{\operatorname{argmin}} \left(\frac{Dist^{c}_{local}}{K_c} \right)$$

The method has been evaluated on several public contemporary datasets of different writing systems and state-of-the-art results are shown to be improved [1].

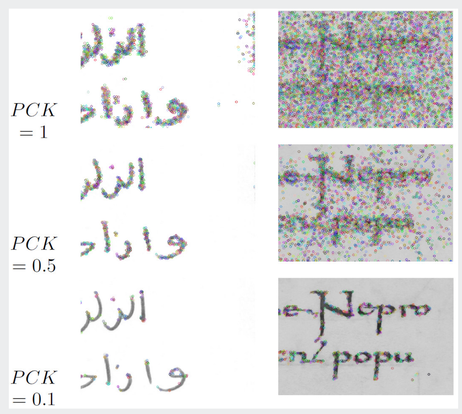
Data Used in the Analysis

100 pages from the "Stiftsbibliothek" library of St. Gall collection [3] were selected for our analysis : 10 scribes, 10 pages per scribe. Sample images are shown below:



Analysing FAST [5] Keypoints Threshold

FAST keypoints were detected with different values of PCK (Percentage of Considered Keypoints). The first column contains part of an image from ICFHR-2016 dataset [6], while the second column contains part of a representative image from St. Gall dataset [3].

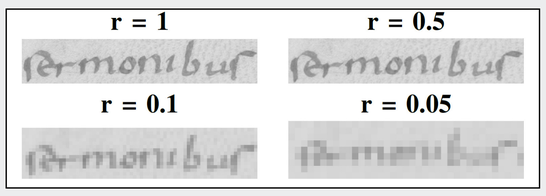


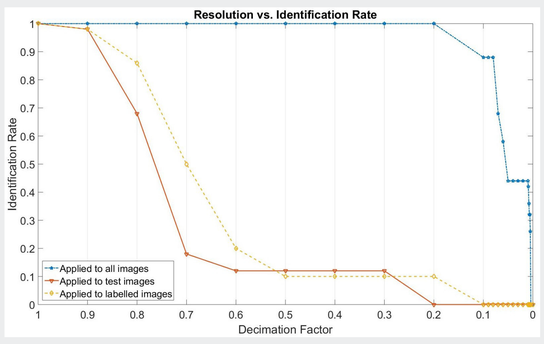
Analysis w.r.t. Degradation

In order to measure the robustness and stability of the method proposed in [1], we analysed it w.r.t. two of the typical degradation types in digital manuscripts, namely resolution and contrast. This selection is based on the prevalence in historical manuscripts and their direct influence on parameter selection of the implemented software tool.

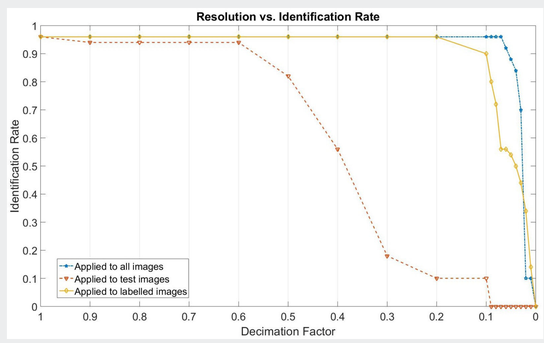
MCT = Maximum Contrast Threshold
r = decimation factor

Resolution



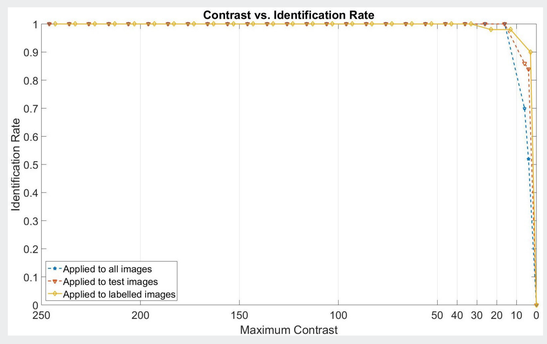


Using FAST Keypoints

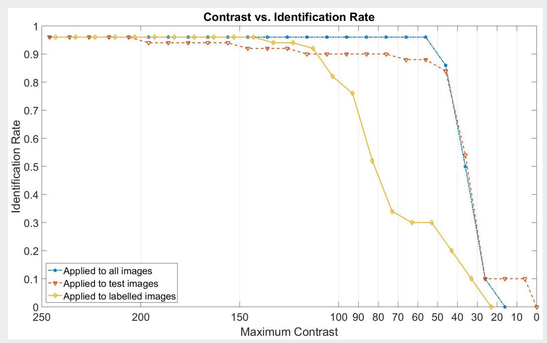


Using SIFT Keypoints

Contrast



Using FAST Keypoints



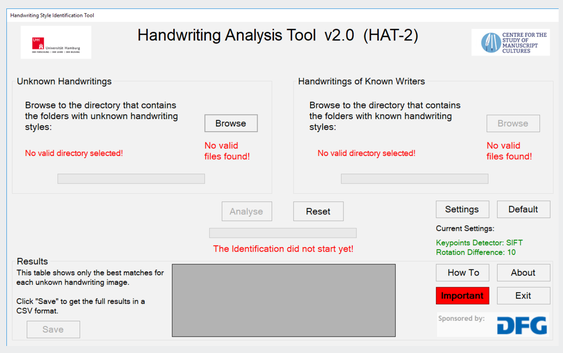
Using SIFT Keypoints

Experimental Results on WI-Historical ICDAR2017 Dataset [7] and a Comparison with Best Results in the Competition:

Method	Top-1	mAP	Dataset details
Normalised LNBNN top 5% FAST keypoints	85.6	68.3	720 writers, 3600 pages 5 pages per writer (mostly English) Leave-one-out
Groningen [18]	76.1	54.2	
Tebessa II [18]	76.4	55.6	
Tebessa I [18]	74.4	52.5	

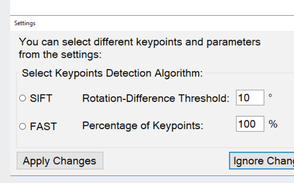
Software Tool Implementation

Our Handwriting Analysis Tool v2.0 (HAT-2) [4] has been made available as open source for analysing handwritings of known scribes and ranking them according to their similarity to unknown handwritings. A quantitative similarity score is computed for each style (scribe) in order to afford the user a relative though rank-based comparison between the styles w.r.t. a given unknown handwriting.



Example of summary results:

	File	Best Match	Score
▸	Unknown1	Fischer	71.3
	Unknown2	Schmidt	80.7
	Unknown3	Schneider	48.1



Example of a Full results file:

	A	B	C
1	Results for Unknown1		
2	Rank	Directory	Score
3	1	Fischer	71.3
4	2	Schneider	15.3
5	3	Schmidt	13.2
6			
7	Results for Unknown2		
8	Rank	Directory	Score
9	1	Schmidt	80.7
10	2	Fischer	10.4
11	3	Schneider	8.7
12			
13	Results for Unknown3		
14	Rank	Directory	Score
15	1	Schneider	48.1
16	2	Fischer	31.8
17	3	Schmidt	19.9

References

[1] H. Mohammed, V. Märgner, T. Konidaris, and H. S. Stiehl, "Normalised local näi ve bayes nearest-neighbour classifier for offline writer identification," in 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, pp. 1013–1018. 2017.

[2] S. McCann and D. G. Lowe, "Local Naïve Bayes Nearest Neighbor for image classification," IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3650–3656, Jun. 2012.

[3] e-codices Virtual Manuscript Library of Switzerland. St. Gallen, stiftsbibliothek. [Online]. Available: http://www.e-codices.ch.

[4] H. Mohammed (2018) Handwriting Analysis Tool v2.0 (HAT-2). [Online]. Available: https://www.manuscript-cultures.unihamburg. de/hat.html.

[5] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection" IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 1, pp. 105–119, 2010.

[6] C. Djeddi, S. Al-Maadeed, A. Gattal, I. Siddiqi, A. Ennaji, and H. El Abed, "ICFHR2016 competition on multi-script writer demographics classification using" QUWI" database."

[7] S. Fiel, F. Kleber, M. Diem, V. Christlein, G. Louloudis, S. Nikos, and B. Gatos, "ICDAR2017 competition on historical document writer identification (historical-wi)," in 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017, pp. 1377–1382.