Exploring Handwritten Document Collections: An EPSC-Based Approach for Feature Extraction and Similarity Analysis

Anders Hast[1,*,†], Örjan Simonsson[2,†]

[1]Uppsala University, Lägerhyddsvägen 1, Uppsala, 751 05, Sweden
[2]Uppsala County Archive on Popular Movement, S:t Olofsgatan 15, Uppsala, 753 21, Sweden

Abstract

This work in progress paper presents a novel approach for the classification and analysis of handwritten documents using a combination of Embedded Prototype Subspace Classification (EPSC) and advanced clustering techniques. We focus on facilitating the examination of document collections by enabling efficient comparisons between documents written by different hands. Our methodology involves the extraction of features from keypoints detected in the handwritten text, which are then processed using t-SNE and modified K-Means clustering to identify clusters of similar features. The novelty lies in a similarity score that is computed to quantify the likeness between document pairs, enabling the identification of stylistic similarities even in the absence of ground truth. An interactive visual application is developed to assist users in exploring the collection, providing insights into the nature of each document, including the differentiation between typewritten and handwritten texts. Our preliminary experiments demonstrate promising results, indicating that documents of the same hand tend to cluster together while distinguishing between varying writing styles. However, we acknowledge that there is room for improvement, particularly in optimising the keypoint detection, feature extraction, and background removal processes, as well as in determining optimal thresholds. Future work will address these limitations, enhancing the robustness of our method and expanding its applicability to a wider range of documents.

Keywords

Document Collections, Writing style, Visualisation, Exploration

1. Introduction

We present a work-in-progress aimed at facilitating a quick and efficient overview of documents with varying handwriting styles, which is crucial for many applications in historical document analysis and archival research. Handwriting analysis has traditionally been a challenging task. However, with the increasing volume of digitized handwritten documents, particularly in historical collections, there is a growing need for automated methods to assist researchers in managing and exploring these collections more efficiently.

IRCDL 2025: 21st Conference on Information and Research Science Connecting to Digital and Library Science, February 20-21 2025, Udine, Italy

[*]Corresponding author.

[†]These authors contributed equally.

anders.hast@it.uu.se (A. Hast); orjan.simonson@fauppsala.se (Simonsson)
https://andershast.com/ (A. Hast); https://https://www.fauppsala.se/ (Simonsson)
0000-0003-1054-2754 (A. Hast); 0000-0002-6347-4102 (Simonsson)

Deep learning methods have demonstrated effectiveness in identifying writing hands but require substantial amounts of training data [1]. In contrast, computer vision-based methods leverage keypoints and local features, which are either clustered to generate a supervector representing each document [2] or processed through a deep learning network [3]. In this paper, we adopt the keypoint-based approach, incorporating several notable modifications as detailed below.

2. Method

Each document, as shown in Figure 1a, is processed as follows: First, the background is removed to enhance the visibility of the text [4], minimising the impact of background noise and ensuring that keypoints are computed only on the text strokes. However, binarisation is not applied here, as it would introduce unwanted jaggedness and distort the text strokes. Instead, the background is removed while preserving the text in grayscale, allowing for accurate keypoint detection without interference from binarisation artifacts. Keypoints are computed using the Harris detector [5], which identifies descriptive points in the script, as shown in Figure 1b.

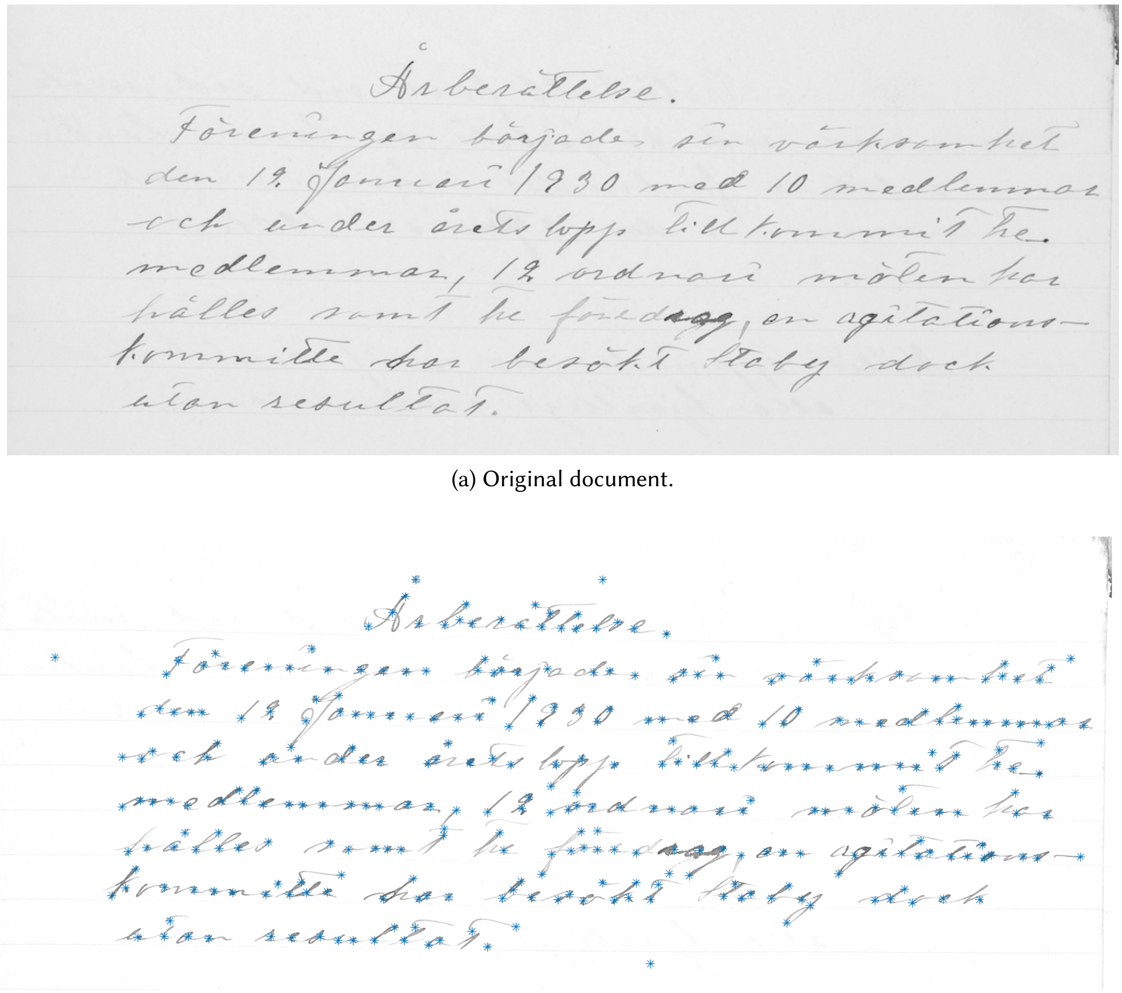
Next, simple yet descriptive feature vectors are computed using the *Radial Line Fourier Descriptor for Historical Handwritten Text Representation* [6]. While more advanced descriptors, such as the SIFT descriptor [7], could have been employed alongside sophisticated keypoints, as shown in [3, 2], these methods have both strengths and limitations. For instance, SIFT offers scale invariance, but its inherent rotation invariance is disadvantageous in this context, where stroke direction is crucial. Additionally, SIFT detects blobs using the difference of Gaussian approach, in contrast to the Harris detector, which focuses on corners. Future work will aim to refine scale invariance to better accommodate documents with varying script sizes.

For each image, the features obtained at each keypoint are processed using t-SNE [8], followed by a modified K-Means clustering algorithm to identify groups of similar features [9], as illustrated in Figure 2. Unlike traditional K-Means, where initial cluster centroids are chosen randomly, the starting values in this method are strategically placed at the cluster centers on a predefined scale. This modification helps improve the clustering by providing a more informed starting point, leading to more accurate grouping of similar features.

While this approach shares similarities with previously published methods, such as keypoint clustering, its novelty lies in using clusters as subspaces that capture variations in stroke execution with similar appearances, as well as in the novel computation of supervectors. The following section outlines how these clusters are employed to develop a classifier.

Recently, *Embedded Prototype Subspace Classification* (EPSC) [10, 9, 11, 12, 9] has been successfully applied to classify diverse datasets. This classifier builds on concepts developed by Kohonen and others [13, 14, 15, 16, 17] and can be viewed as a two-layer neural network [11, 17, 18], where the weights are mathematically derived using Principal Component Analysis (PCA) [18]. Prototypes are identified using t-SNE and the modified K-Means clustering, as previously described, where each feature within a cluster serves as a prototype. This approach makes the process easy to interpret through visualisations.

To compare two documents, **A** and **B**, we follow this procedure: The first eigenvector from the PCA of each cluster in document **A** is projected into the subspaces obtained from the



Årberättelse.
Föreningen började sin värksamhet
den 19. Januari 1930 med 10 medlemmar
och under årets lopp tillkommit 7 ne.
medlemmar, 12 ordinari möten har
hålles samt 1e föredrag, en agitations-
kommitté har besökt Stoby dock
utan resultat.

(a) Original document.

(b) The background has been removed and keypoints appear on the text strokes.

Figure 1: The documents (a) are processed so that the background is removed and keypoints are computed (b).

clusters of document **B**. Notably, using only the first eigenvector in the subspace appears to yield reasonable results. Therefore, the comparison primarily relies on the dot product of the first eigenvectors, as the projection depth is effectively 1. Future work will explore the impact of varying projection depths to further refine this approach.

To track the similarity between clusters in documents **A** and **B**, we compute two lists. The first list contains the identities of the clusters in **A**, while the second list, denoted as Act, holds the corresponding activation values. Given that two documents typically have differing numbers of clusters, and that multiple clusters in **B** can correspond to a single cluster in **A**, we define the similarity between the documents as the ratio of activated clusters in **A** that exceed a certain activation threshold θ to the total number of activated clusters in **A**.

The similarity ratio R is calculated as follows:

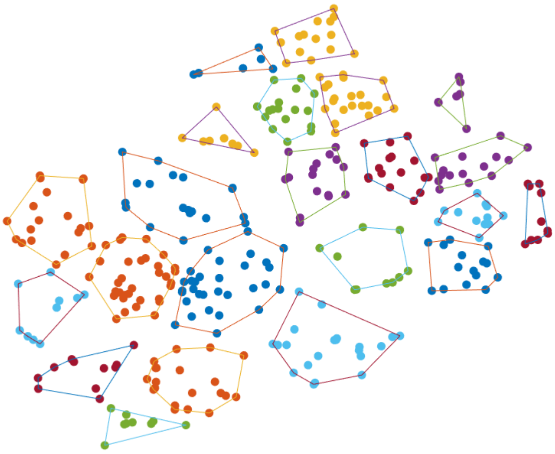


Figure 2: The features extracted at each keypoint from a document image are processed using t-SNE, followed by modified K-Means clustering to identify clusters of similar features.

$$R = \frac{\sum_{i=1}^{n} \text{ if } Act[i] > \theta}{\sum_{i=1}^{n} \text{ if } Act[i] \neq 0} \tag{1}$$

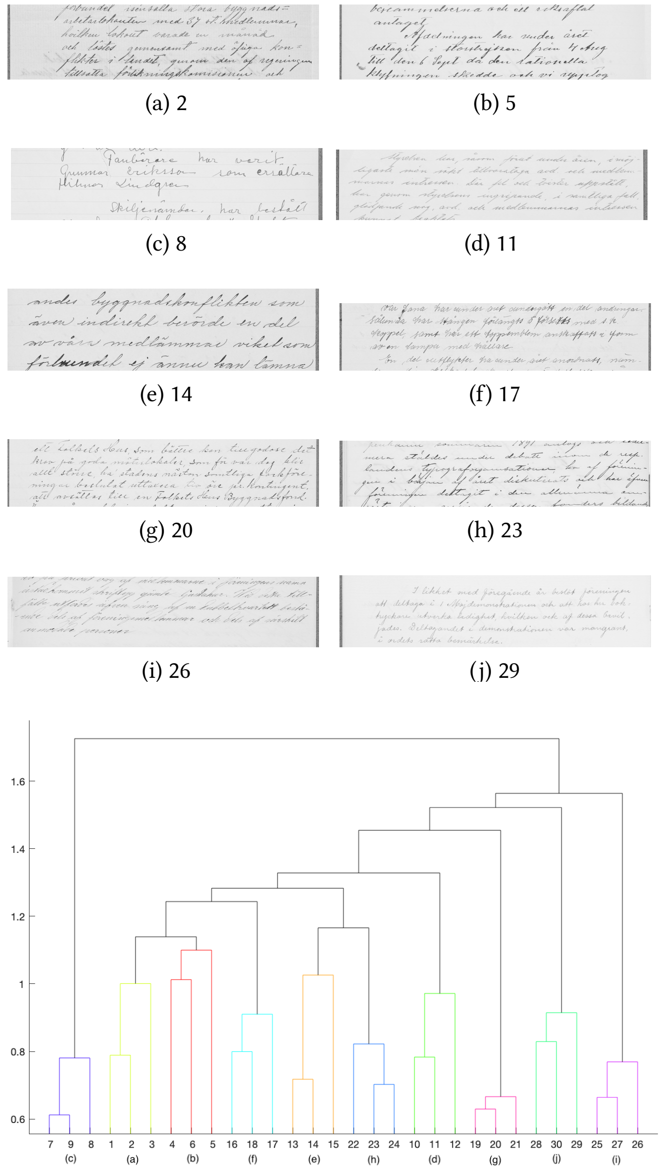
Here, n represents the number of elements in the list Act. In our experiments, we set $\theta = 0.92$. However, further experimentation is necessary to determine the optimal threshold, as it may vary depending on the specific methods used for feature extraction and clustering.

One advantage of the approach presented above is that each document is uniquely characterised by its own description based on EPSC. However, the supervector is derived from a row in the correlation matrix, where each document page is compared against all other pages. The key idea is that pages written by the same hand are expected to show higher similarity, reflected in a greater R value, whereas pages written by different hands should exhibit lower similarity, resulting in a smaller R value.

3. The Labour's memory Project

While the primary objective of the approach outlined here is not writer identification but rather to facilitate the browsing of document collections, it can still be effectively utilised for that purpose. Both tasks are conducted using data from the Labour's Memory infrastructure project, launched in 2021 to digitise and present annual reports and financial records from blue-collar labour organisations spanning the period from 1880 to 2020. The project includes materials from Swedish unions at various levels (local, district, and national) and international labor organisations, spanning repositories such as *Folkrörelsearkivet för Uppsala län* (FAC) and *Arbetarrörelsens arkiv och bibliotek* (ARAB). The corpus, primarily in Swedish with some English, German, French, and Spanish, is estimated to consist of 1–1.5 million pages, with 300,000 pages digitised by 2024.

The local organizations' annual reports, housed at FAC, consist of approximately 35,000 pages, often handwritten or typewritten and rarely professionally published. These texts,



(a) 2

(b) 5

(c) 8

(d) 11

(e) 14

(f) 17

(g) 20

(h) 23

(i) 26

(j) 29

(k) Dendrogram, showing how the 10 documents with three pages each are perfectly classified.

Figure 3: Figures a-j show examples from each document group as shown in figure (k). The numbers correspond to the number in figure 3k.

along with their manually transcribed counterparts, are essential for developing handwriting text recognition (HTR) models. In contrast, national organisations produced professionally printed annual reports for broader audiences. The handwritten reports, often created by secretaries, chairs, auditors, or cashiers, exhibit diverse styles, reflecting varying levels of skill and consistency.

3.1. Writer Identification

An experiment was conducted in which three pages from ten different documents, each written by distinct authors, were selected. The EPSC for each document was computed, and the similarity score, denoted as R, was calculated as described herein. To make the experiment more challenging, the last page of each document was also included, where often only half of the pages contained written text.

In figure 3, one page from each of the ten documents are shown, together with a dendrogram that was computed from a correlation matrix, which was computed by all similarity scores R, as previously explained. Since there is no ground truth available, the documents had to be selected manually, and therefore we chose to use only a limited number. Although the grouping of documents achieved perfect results, this is not always guaranteed. Nevertheless, the writing styles of some documents are quite similar, which provides an indication of the performance of the proposed algorithm.

3.2. Browsing the collection

An interactive application was developed to enable users to browse a subsection of the entire document collection, specifically consisting of the already transcribed documents. Each document is represented as a blue dot in the t-SNE visualization in figure 4. The document selected by the user is displayed in the lower left corner, while a close-up view is shown in the lower right corner, enabling a more detailed examination of the writing style. The identifying number and name of the chosen document is shown in the top.

Interestingly, it was discovered that there were numerous typewritten documents in the collection. The user was able to identify them within just a few minutes, as they are located to the left of the red curve, which was added to highlight their position. Browsing through all 1,700 documents manually, on the other hand, would have been rather cumbersome.

4. Conclusion and Future work

The two small experiments demonstrate promising results, suggesting that the proposed approach is effective. However, there is significant room for improvement. As previously mentioned, both the keypoint detector and the feature extractor can be optimised, and the threshold θ should be further investigated. The clustering shown in Figure 4 indicates that improvements ought to be possible, as typewritten documents should ideally occupy a distinct region separate from handwritten documents. However, it is important to note that some documents contain areas with both typewritten and handwritten text. Furthermore, the background removal process could be optimised to effectively handle documents with varying levels of background noise.

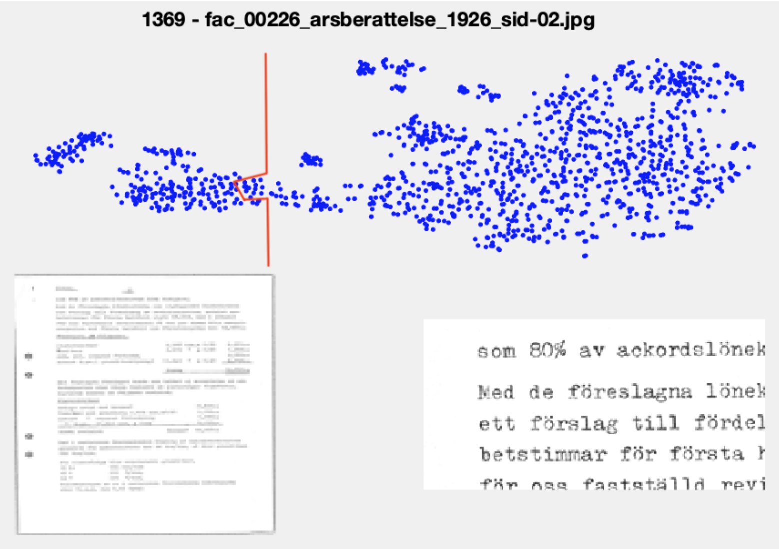


Figure 4: Thanks to the visual interactive application, it was possible to quickly identify which documents were typewritten and which were not, all within a few minutes. The typewritten documents are located to the left of the red line, which was added later.

Acknowledgments

This work was partially funded through the *Labour's Memory* project, which received support from Riksbankens Jubileumsfond under Grant Agreement No. IN20-0040.

References

[1] M. Kestemont, V. Christlein, D. Stutzmann, Artificial paleography: Computational approaches to identifying script types in medieval manuscripts, Speculum 92 (2017) S86–S109. URL: https://doi.org/10.1086/694112. doi:10.1086/694112. arXiv:https://doi.org/10.1086/694112.

[2] S. Fiel, R. Sablatnig, Writer identification and writer retrieval using the fisher vector on visual vocabularies, in: 2013 12th International Conference on Document Analysis and Recognition, 2013, pp. 545–549. doi:10.1109/ICDAR.2013.114.

[3] V. Christlein, M. Gropp, S. Fiel, A. Maier, Unsupervised Feature Learning for Writer Identification and Writer Retrieval, in: IEEE (Ed.), 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 991–997. URL: https://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2017/Christlein17-UFL.pdf. doi:10.1109/ICDAR.2017.165.

[4] E. Vats, A. Hast, P. Singh, Automatic document image binarization using bayesian optimization, in: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing, ACM, 2017, pp. 89–94.

[5] C. Harris, M. Stephens, A combined corner and edge detector., in: Proceedings of The Fourth Alvey Vision Conference, Manchester, UK, 1988, pp. 147–151.

[6] A. Hast, E. Vats, Radial line fourier descriptor for historical handwritten text representation, Journal of WSCG 26 (2018) 31–40. doi:10.24132/JWSCG.2018.26.1.4.

[7] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2004) 91–110.

[8] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (2008) 2579–2605.

[9] A. Hast, E. Vats, Word recognition using embedded prototype subspace classifiers on a new imbalanced dataset., Journal of WSCG 29 (2021) 39–47. URL: http://wscg.zcu.cz/WSCG2021/2021-J-WSCG-1-2.pdf.

[10] A. Hast, Magnitude of semicircle tiles in fourier-space : A handcrafted feature descriptor for word recognition using embedded prototype subspace classifiers, Journal of WSCG 30 (2022) 82–90. doi:10.24132/JWSCG.2022.10.

[11] A. Hast, M. Lind, E. Vats, Embedded prototype subspace classification : A subspace learning framework, in: The 18th International Conference on Computer Analysis of Images and Patterns (CAIP), Lecture Notes in Computer Science, 2019, pp. 581–592.

[12] A. Hast, M. Lind, Ensembles and cascading of embedded prototype subspace classifiers, Journal of WSCG 28 (2020) 89–95. doi:10.24132/JWSCG.2020.28.11.

[13] W. Watanabe, P. F. Lambert, C. A. Kulikowski, J. L. Buxto, R. Walker, Evaluation and selection of variables in pattern recognition, in: J. Tou (Ed.), Computer and Information Sciences, volume 2, New York: Academic Press, 1967, pp. 91–122.

[14] T. Kohonen, P. Lehtiö, J. Rovamo, J. Hyvärinen, K. Bry, L. Vainio, A principle of neural associative memory, Neuroscience 2 (1977) 1065 – 1076. URL: http://www.sciencedirect.com/science/article/pii/0306452277901294. doi:https://doi.org/10.1016/0306-4522(77)90129-4.

[15] T. Kohonen, E. Oja, Fast adaptive formation of orthogonalizing filters and associative memory in recurrent networks of neuron-like elements, Biological Cybernetics 21 (1976) 85–95. URL: https://doi.org/10.1007/BF01259390. doi:10.1007/BF01259390.

[16] T. Kohonen, E. Reuhkala, K. Mäkisara, L. Vainio, Associative recall of images, Biological Cybernetics 22 (1976) 159–168. URL: https://doi.org/10.1007/BF00365526. doi:10.1007/BF00365526.

[17] E. Oja, T. Kohonen, The subspace learning algorithm as a formalism for pattern recognition and neural networks, in: IEEE 1988 International Conference on Neural Networks, volume 1, 1988, pp. 277–284. doi:10.1109/ICNN.1988.23858.

[18] J. Laaksonen, Subspace classifiers in recognition of handwritten digits, G4 monografiaväitöskirja, Helsinki University of Technology, 1997-05-07. URL: http://urn.fi/urn:nbn:fi:tkk-001249.